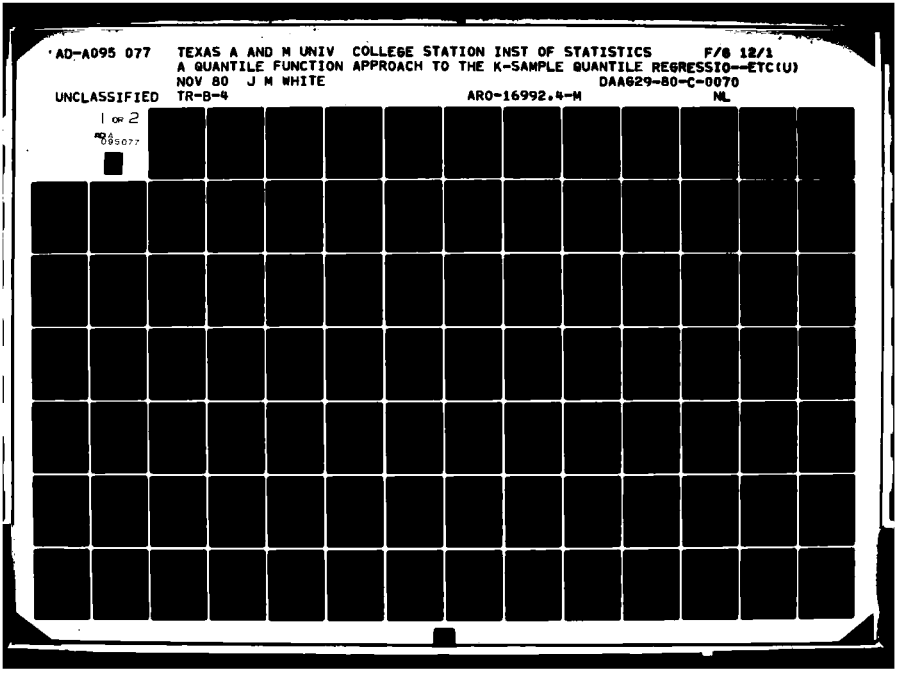

AD-A095 077 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
A QUANTILE FUNCTION APPROACH TO THE K-SAMPLE QUANTILE REGRESSIO--ETC(U)
NOV 80 J M WHITE DAAG29-80-C-0070
UNCLASSIFIED TR-B-4 ARO-16992.4-M NL
1 OF 2
AD A095077

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843

INSTITUTE OF STATISTICS
Phone 713 - 845-3141

A QUANTILE FUNCTION APPROACH TO THE K-SAMPLE QUANTILE REGRESSION PROBLEM

James Michael White

Institute of Statistics, Texas A&M University

Technical Report No. B-4

November 1980

Texas A & M Research Foundation
Project No. 4226

"Robust Statistical Data Analysis and Modeling"

Sponsored by the U.S. Army Research Office
Grant DAAG29-80-C-0070

Professor Emanuel Parzen, Principal Investigator

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report B-4	2. GOVT ACCESSION NO. AD-A094 017	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Quantile Function Approach to the K-Sample Quantile Regression Problem		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) James Michael White		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0070
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Institute of Statistics College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Research Office		12. REPORT DATE November 1980 13. NUMBER OF PAGES 117
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Quantile Function, Quantile Regression, Location-Scale Parameter-estimation, K-sample Comparisons.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See attached sheet.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract

In this diesertation a procedure for estimating the parameters of a quantile regression function is investigated. The procedure is based on the work of Parzen (1979a) in the theory of quantile functions and is applicable to a wide range of distributional families.

The procedure assumes the quantile functions of k populations to be location-scale shifts of a common quantile function. First, a goodness-of-fit procedure for determining the common distributional shape of the k populations generalizes the one-population data modeling techniques of Parzen (1979a). An estimator of the shape parameter of a distribution is also investigated. The methods of Ogawa (1951) and Eubank (1979) are then used for estimating the location and scale parameters of the k populations. A regression model for the location and scale parameters is specified, and the resulting estimators of the regression parameters are used to determine a regression function for any quantiles of the observed data. Finally it is shown that inferences about the quantile relationships can be based on the asymptotic normality of the estimated parameters. The procedures are applied to some published data sets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

1. THE PROBLEM OF K-SAMPLE QUANTILE REGRESSION

The technique of regression analysis is used to model the relationship between the mean of a response variable Y and a predictor variable X. In some situations it may be more useful to model the relationship between the percentiles (or quantiles) of a response variable Y and the values of a predictor variable X.

Hogg (1975), Griffiths and Willcox (1978), and Angers (1979) investigate the relationship between several percentiles of salary level for professors at a major university as a function of their years in service. Hogg (1975) uses a nonparametric graphic technique to estimate linear percentile relationships. Griffiths and Willcox (1978) use a maximum likelihood approach based on assuming the data to have a normal distribution. Angers (1979) adopts a nonparametric approach using linear grafted polynomials. He assumes that a specific dependent relationship exists among the various percentile regression curves.

Reliability and survival analyses often lead to situations where one is interested in modeling percentiles of the survival distribution as a function of the treatment, e.g. modeling the median survival time of fish as a function of water temperature. Matis and Wehrly (1979) investigate the resistance of the green sunfish, Lepomis cyanellus, to various levels of thermal pollution

---

Citations follow the format of the Journal of the American Statistical Association.

using a compartmental models approach. The data consist of survival times of fish at fixed temperatures. They assume the data to have a three-parameter Weibull distribution and estimate all three parameters for several temperatures. LaRiccia (1979) analyzes the same data using minimum quantile distance estimators of the parameters. Our goal is to estimate the relationship between the percentiles of the survival times and the test temperature.

Thus we consider the following statistical situation. Consider random samples from $k(k \geq 2)$ populations, i.e. for $i = 1, \ldots, k$, let $\{Y_{i1}, \ldots, Y_{in_i}\}$ be n_i independent observations of a random variable Y_i which has cumulative distribution function

$$F_i(y) = Pr(Y_i \leq y)$$

and quantile function

$$Q_i(u) = F_i^{-1}(u) = \inf \{y : F_i(y) \geq u\}, \ 0 \leq u \leq 1 \ .$$

Associated with Y_i is a numerical characteristic, X_i, of the ith population and we assume for convenience that $X_1 \leq \ldots \leq X_k$. Thus $X_1, \ldots, X_k$ would be the various years in service or water temperatures in the examples cited above.

The k sample quantile regression problem is to find estimators of and make inferences about $A(u)$ and $B(u)$ in the k-sample regression model

$$Q_i(u) = A(u) + B(u) \ X_i \ , \ i = 1, \ldots, k \ .$$

The purpose of this dissertation is to investigate a method for determining such estimators based on the approach to quantile functions presented by Parzen (1979a).

We assume a location-scale shift model for Q_i, i.e. that $Q_1, \ldots, Q_k$ can be written as a location-scale shift of some common quantile function. We write

$$Q_i(u) = \mu_i + \sigma_i Q_o(u) \quad , i = 1, \ldots, k ,$$

where μ_i and σ_i are the location and scale parameters respectively of Q_i and where either the form of Q_o is unknown but does not depend on any unknown parameters or $Q_o(u) = [Q_o^*(u)]^\gamma$ where $Q_o^*(\cdot)$ is a known, completely specified quantile function and γ is an unknown shape parameter. For example, we may believe that Q_o corresponds to either a standard normal so that

$$Q_o(u) = \Phi^{-1}(u) ,$$

where

$$\Phi(y) = \int_{-\infty}^{y} (1/\sqrt{2\pi}) \exp(-t^2/2)dt$$

or a standard lognormal distribution so that

$$Q_o(u) = \exp(\Phi^{-1}(u)) \quad .$$

On the other hand we may believe that Q_o corresponds to a three-parameter Weibull distribution so that

$$Q_o(u) = (-\log(1 - u))^\gamma$$

where the shape parameter γ needs to be estimated.

We further assume that μ_i and σ_i are linearly related to X_i, i.e.

$$\mu_i = \alpha_\mu + \beta_\mu X_i$$

$$\sigma_i = \alpha_\sigma + \beta_\sigma X_i \quad , i = 1, \ldots, k \quad .$$

Thus, we can write the quantile regression model

$$Q_i(u) = [\alpha_\mu + \beta_\mu X_i] + [\alpha_\sigma + \beta_\sigma X_i] Q_o(u)$$

$$= [\alpha_\mu + \alpha_\sigma Q_o(u)] + [\beta_\mu + \beta_\sigma Q_o(u)] X_i \quad .$$

The aim of this dissertation is to investigate

1) *methods for identifying* the shape of Q_o, i.e. either choose a completely specified function from possible contenders or estimate γ,
2) *methods* for determining estimators $\hat{\alpha}_\mu$, $\hat{\beta}_\mu$, $\hat{\alpha}_\sigma$, $\hat{\beta}_\sigma$, of α_μ, β_μ, α_σ, β_σ, and
3) the properties of estimating A(u) and B(u) by

$$\hat{A}(u) = \hat{\alpha}_\mu + \hat{\alpha}_\sigma Q_o(u) \quad ,$$

$$\hat{B}(u) = \hat{\beta}_\mu + \hat{\beta}_\sigma Q_o(u) \quad .$$

Section 2 presents basic definitions and theorems regarding the quantile function and the empirical quantile function. The Weibull distribution and its properties are also discussed.

In Section 3, Parzen's (1979a) nonparametric data modeling method of determining Q_0 for one population is described and extended to determining a common Q_0 for k populations. An estimator of the shape parameter γ is proposed and its properties are investigated.

In Section 4 we discuss two formulations for estimating location and scale parameters using linear combinations of order statistics. The approaches are due to Ogawa (1951) and Eubank (1979).

In Section 5 we develop new methods for k-sample quantile regression using the models discussed above. Hypothesis testing procedures are provided. The application of the technique to a particular type of location-scale comparison problem is also discussed.

Finally in Section 6 the techniques of Sections 3 through 5 are applied to the analysis of the Hogg data and the Matis and Wehrly data.

Section 7 consists of conclusions and suggested topics for future research.

2. QUANTILE FUNCTIONS AND THE WEIBULL DISTRIBUTION

In Section 2.1 we introduce the quantile function notation of Parzen (1979a) and state some useful theorems and properties of the quantile function. In Section 2.2 we define the Weibull and extreme value distributions and provide plots of the Weibull quantile and density quantile functions for a range of values of its shape parameter. Lower bounds on the variance for unbiased estimators of the parameters of the Weibull distribution are given.

2.1 Definitions and Notation of the Quantile Function Approach

We adopt the quantile function notation of Parzen (1979a). Some useful definitions are:

1. The cumulative distribution function (cdf) of a random variable X is defined by $F(x) = \Pr(X \leq x)$.
2. The quantile function of X, Q(u), is defined by

$$Q(u) = F^{-1}(u) = \inf\{x: F(x) \geq u\}, \quad 0 \leq u \leq 1 .$$

3. The probability density function of a continuous random variable X is defined by

$$f(x) = d\,F(x) \,/\, dx$$

so that

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad .$$

4. The quantile density function, q(u), is defined by

$$q(u) = dQ(u) / du \ , \ 0 < u < 1 \ .$$

5. The density quantile function, fQ(u), is defined by

$$fQ(u) = f(Q(u)) \ , \ 0 \leq u \leq 1 \ .$$

The sample analogs of the above quantities are presented in the following definitions. Let $X_{1;n} \leq \ldots \leq X_{n;n}$ be the order statistics of a random sample of size n from a population with cdf F.

6. The empirical distribution function (edf), $\tilde{F}(x)$, is given by

$$\tilde{F}(x) = 0 \quad \text{if } x < X_{1;n}$$

$$= j/n \quad \text{if } X_{j;n} \leq x < X_{j+1;n} \ ,$$

$$j = 1, \ldots, n-1$$

$$= 1 \quad \text{if } X_{n;n} \leq x$$

or

$$\tilde{F}(x) = 1/n \sum_{j=1}^{n} \delta_{X_j}(x) \ ,$$

where

$$\delta_X(x) = 1 \quad \text{if } X \leq x$$

$$= 0 \quad \text{otherwise} \ .$$

7. The empirical quantile function, $\tilde{Q}(u)$, is defined by

$$\tilde{Q}(u) = \tilde{F}^{-1}(u)$$

$$= X_{j;n} \ , \ (j-1)/n < u \leq j/n \ ,$$

$$j = 1, \ldots, n \ .$$

While this is a natural definition of $\tilde{Q}(u)$, two other continuous definitions discussed by Parzen (1979a) are useful in both theoretical and applied problems. The piecewise linear version of $\tilde{Q}(u)$ is defined by

$$\tilde{Q}_L(u) = n[j/n-u]X_{j-1;n} + n[u - (j-1)/n] \ X_{j;n} \ ,$$

$$(j-1)/n \leq u \leq j/n \ ,$$

$$j = 1, \ldots, n \ , \qquad (2.1.1)$$

where $X_{0;n}$ is a natural minimum, i.e. a lower bound on the range of the data, if one exists, and $X_{0;n} = X_{1;n}$ otherwise.

The shifted piecewise linear version of $\tilde{Q}(u)$ is defined by

$$\tilde{Q}_S(u) = n[(j + .5)/n - u]X_{j;n} + n[u - (j - .5)/n]X_{j+1:n} \ ,$$

$$(j - .5)/n \leq u \leq (j + .5)/n,$$

$$j = 1, \ldots, n - 1 \ .$$

We leave $\tilde{Q}_S(u)$ undefined for $u < .5/n$ or $u > 1 - .5/n$.

8. If we use $\tilde{Q}_S(u)$ then we can define the empirical quantile density, $\tilde{q}(u)$, as

$$\tilde{q}(u) = d\tilde{Q}_S(u)/du$$

$$= n(X_{j+1;n} - X_{j;n}),\ (j-.5)/n < u < (j+.5)/n,$$

$$j = 1, \ldots, n-1, \qquad (2.1.2)$$

and the empirical density quantile function, $\widetilde{fQ}(r)$, as

$$\widetilde{fQ}(u) = 1 / \tilde{q}(u) \quad .$$

Two useful properties of the quantile function are given in Theorems 2.1.1 and 2.1.2.

Theorem 2.1.1: Let $F(\cdot)$ be a strictly increasing cdf and let $g(\cdot)$ be a strictly increasing continuous function. If $Y = g(X)$, then $Q_Y(u) = g(Q_X(u))$. If $g(\cdot)$ is strictly decreasing, then $Q_Y(u) = g(Q_X(1-u))$.

Thus, if $Y = \mu + \sigma X$, then $Q_Y(u) = \mu + \sigma Q_X(u)$, and if $Y = \log(X)$, then $Q_Y(u) = \log(Q_X(u))$. This property of the quantile function provides a natural representation for parameter estimation since it allows one to formulate the estimation of location and scale parameters as the estimation of parameters in the simple linear regression of Q_Y on Q_X if Q_X is a simple known function.

Theorem 2.1.2: Let $fQ(\cdot)$ and $q(\cdot)$ be the density quantile and quantile density functions corresponding to $Q(\cdot)$. Then $fQ(u) = 1/q(u)$.

Definitions and useful theorems regarding the asymptotic distribution of $\tilde{Q}(u)$ follow.

9. A Brownian Bridge process $\{B(u),\ 0 \leq u \leq 1\}$ is a zero mean normal process with covariance kernel

$$K_B(u_1, u_2) = \mathrm{Cov}(B(u_1), B(u_2)) = \min(u_1, u_2) - u_1 u_2 \ .$$

Theorem 2.1.3: Under suitable conditions (see Csörgö and Révész 1978)

$$\sqrt{n}\ fQ(u)(\tilde{Q}(u) - Q(u)) \xrightarrow{d} B(u), \text{ for all } u\ ,$$

where the symbol $\xrightarrow{d}$ denotes "converges in distribution to".

A special case of this convergence theorem is the following:

Theorem 2.1.4: Let F be an absolutely continuous cdf with pdf f and let $0 < u_1 < \ldots < u_r < 1$. If fQ is differentiable in a neighborhood of u_i and $fQ(u_i) \neq 0$, for all i, then

$$\sqrt{n}\ (\tilde{\underset{\sim}{Q}} - \underset{\sim}{Q}) \xrightarrow{d} N_r(\underset{\sim}{0}_r, C)$$

where

$$\tilde{\underset{\sim}{Q}} = (\tilde{Q}(u_1), \ldots, \tilde{Q}(u_r))^T,$$

$$\underset{\sim}{Q} = (Q(u_1), \ldots, Q(u_r))^T,$$

$$\underset{\sim}{0}_r = (0, \ldots, 0)^T,$$

and

$$C = (C_{ij})$$

where

$$C_{ij} = C_{ji} = \frac{u_i(1 - u_j)}{fQ(u_i)fQ(u_j)}, \quad 1 \leq i \leq j \leq r \quad . \tag{2.1.3}$$

2.2 The Weibull and Extreme Value Distributions

In this dissertation the basic model for Q(u) is to assume

$$Q(u) = \mu + \sigma Q_o(u)$$

where $Q_o(u)$ is a completely specified quantile function except for a possibly unknown shape parameter and μ and σ are the location and scale parameters of Q. Two quantile functions that have proven to be particularly useful in a variety of statistical problems are those of the three-parameter Weibull distribution and the extreme value distribution. The Weibull and extreme value distributions have been used as models in reliability, survival studies, quality control, hydrology, etc. (see Dubey 1967; Hassanein 1971; Johnson and Kotz 1970).

Definition: A continuous random variable Y is said to have the three-parameter Weibull distribution with parameters μ, σ, and c if

$$\begin{aligned} F(y) &= 0 \quad \text{if } y \leq \mu \\ &= 1 - \exp\{-[(y - \mu)/\sigma]^c\} \quad \text{if } y > \mu \end{aligned} \tag{2.2.1}$$

where σ, c are greater than zero.

The parameters μ, σ, and $c(= 1/\gamma)$ are the location, scale, and shape parameters, respectively. For a random variable following a three-parameter Weibull distribution we have

$$Q(u) = \mu + \sigma[-\log(1-u)]^{\gamma},\ \gamma = \frac{1}{c},\ 0 \leq u \leq 1,$$

$$Q_o(u) = [-\log(1-u)]^{\gamma},$$

$$f_o(y) = c\, y^{c-1} \exp(-y^c),$$

and

$$f_o Q_o(u) = (1/\gamma)(1-u)[-\log(1-u)]^{1-\gamma}.$$

By varying the shape parameter γ, one can fit a wide range of unimodal distributional shapes from skewed right to almost symmetric to skewed left. The role of μ is as a threshhold value (or starting value), i.e. $Q(0) = \mu$, rather than as a measure of central tendency. Figures 2.A to 2.J display the Weibull Q_o and $f_o Q_o$ functions for $\gamma = .1(.1)1$.

By letting $c = 1$(or $\gamma = 1$) in 2.2.1 we obtain the exponential distribution. For $c < 3$(or $\gamma > .333$) the distribution is skewed right. When $3 < c < 4$ (or $.25 < \gamma < .333$) the distribution looks more symmetric. For $c = 3.6$ the Weibull density is similar to that of the normal giving $\sqrt{b_1} = .0006$ and $b_2 = 2.7167$ (Kübler 1979) where $\sqrt{b_1}$ is the skewness measure and b_2 measures kurtosis (see Rao 1973, p. 101). For $c > 4$(or $\gamma < .25$) the distribution is skewed left.

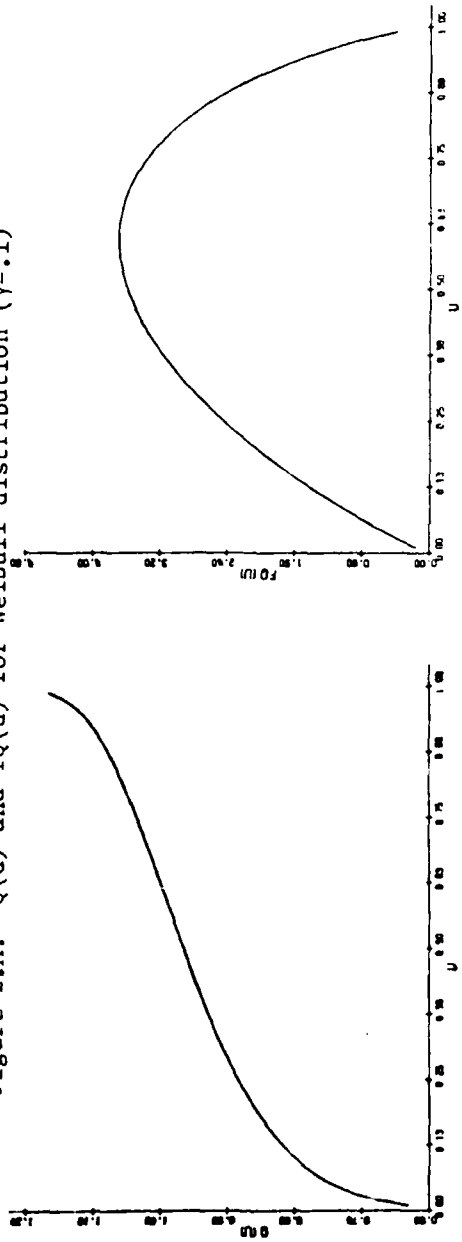


Figure 2.A: Q(u) and fQ(u) for Weibull distribution (γ=.1)

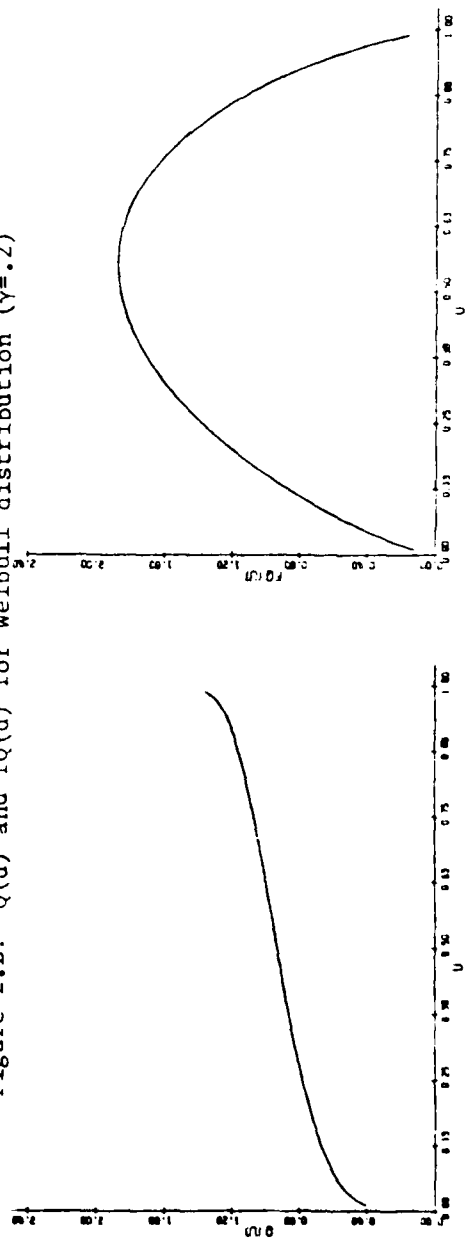


Figure 2.B: Q(u) and fQ(u) for Weibull distribution (γ=.2)

Figure 2.C: Q(u) and fQ(u) for Weibull distribution (γ=.3)

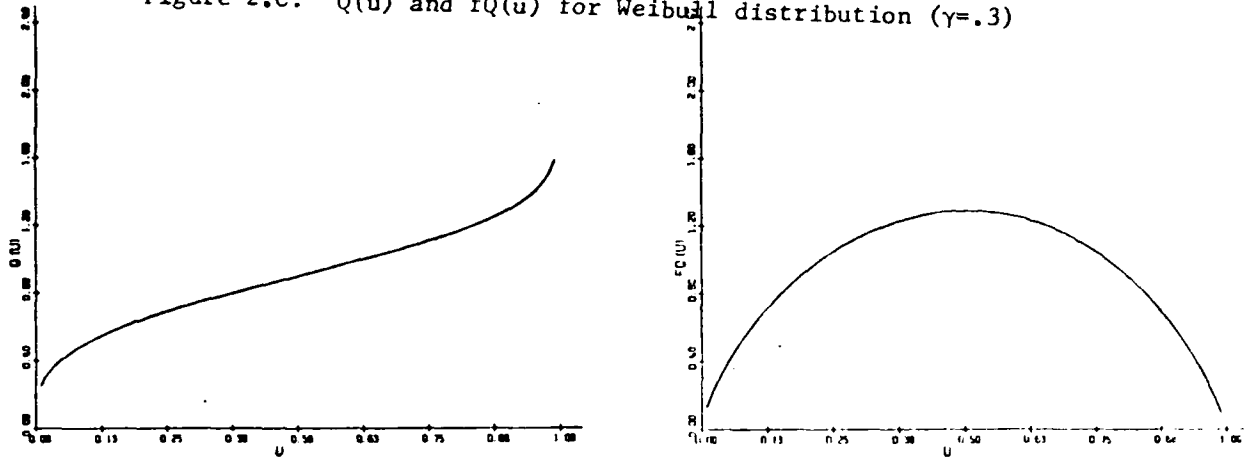


Figure 2.D: Q(u) and fQ(u) for Weibull distribution (γ=.4)

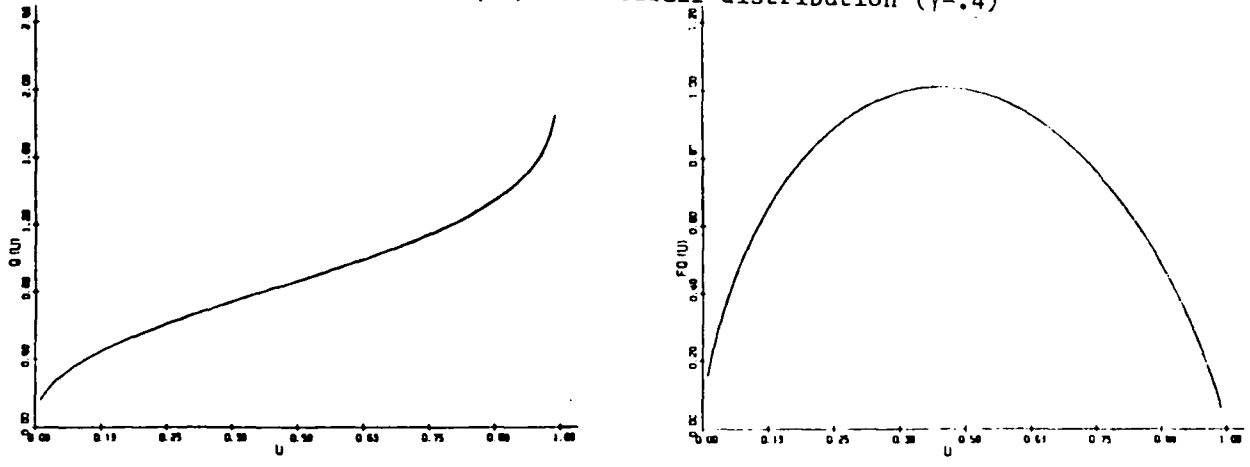


Figure 2.E: Q(u) and fQ(u) for Weibull distribution (γ=.5)

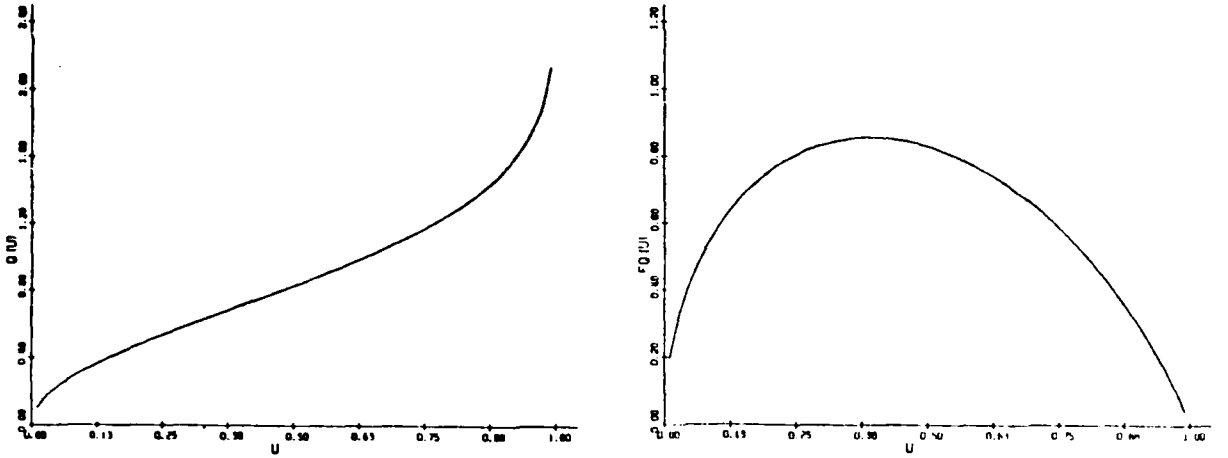
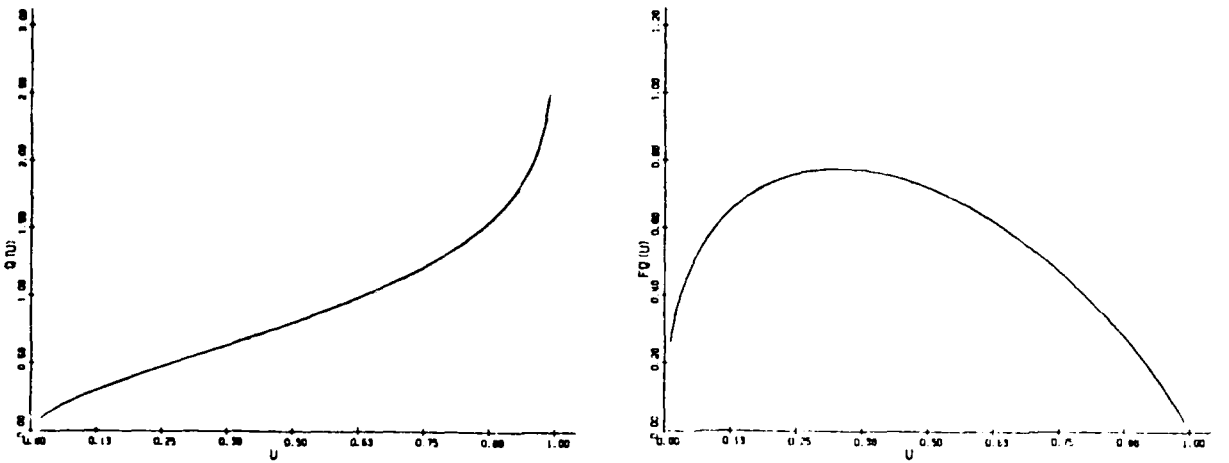


Figure 2.F: Q(u) and fQ(u) for Weibull distribution (γ=.6)



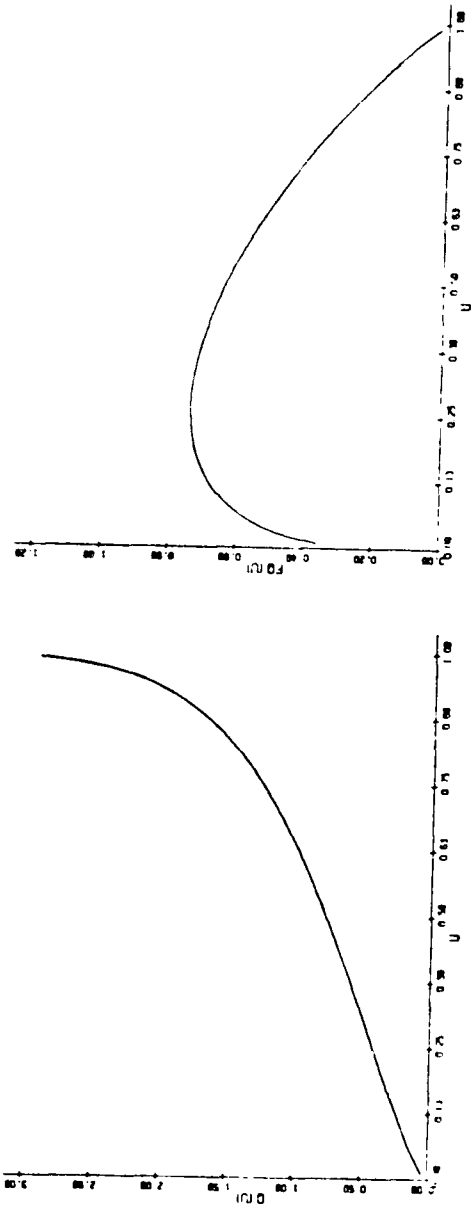


Figure 2.G: Q(u) and fQ(u) for Weibull distribution (γ=.7)

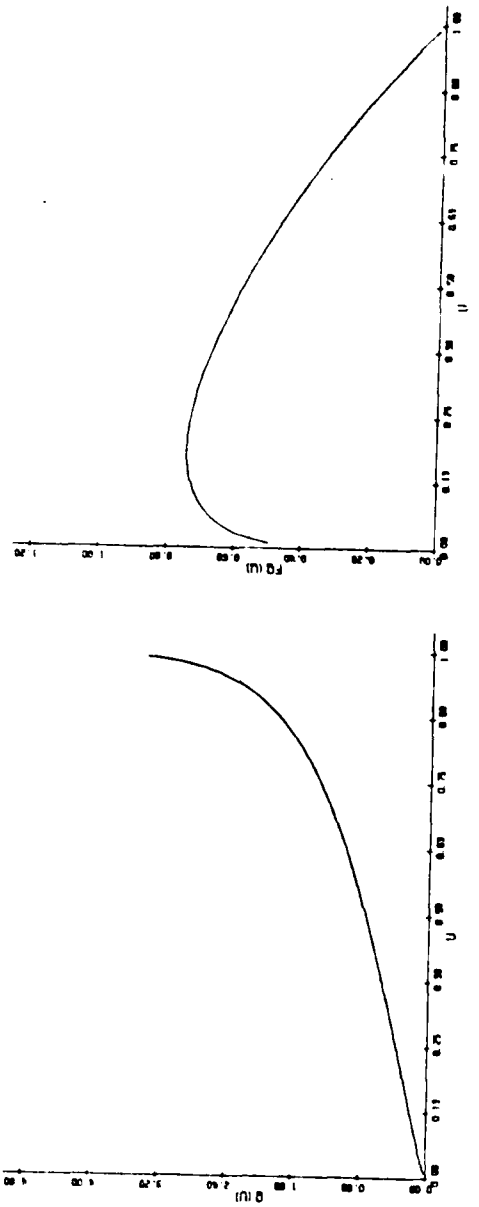


Figure 2.H: Q(u) and fQ(u) for Weibull distribution (γ=.8)

Figure 2.I: Q(u) and fQ(u) for Weibull distribution (γ=.9)

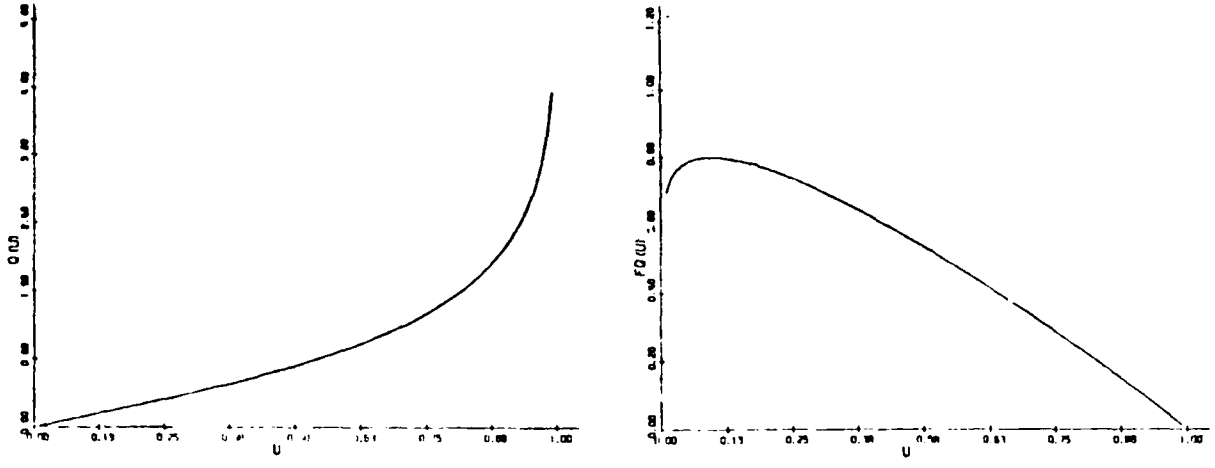
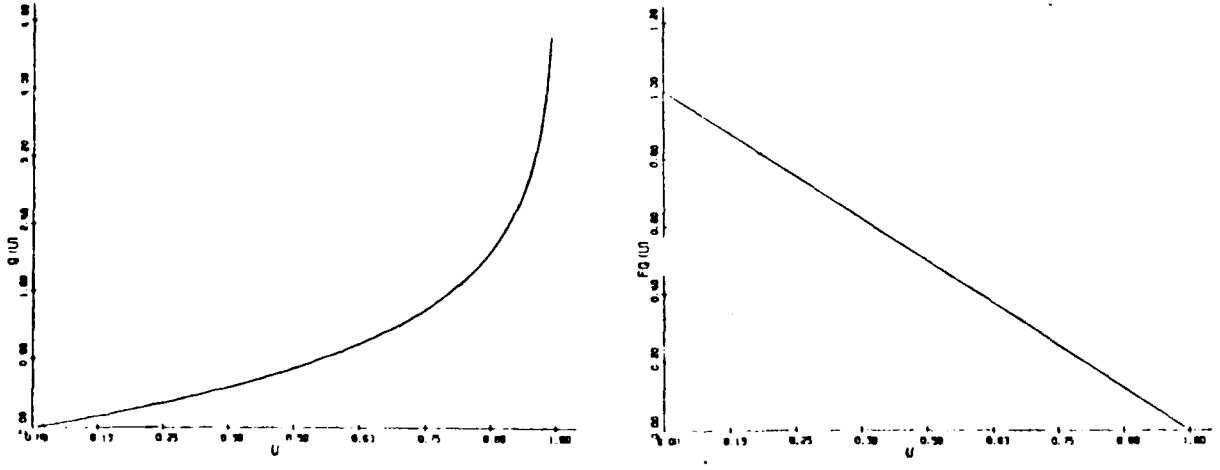


Figure 2.J: Q(u) and fQ(u) for Weibull distribution (γ=1.0)



Definition: If the continuous random variable Y has the three-parameter Weibull distribution, then $X = -\log(Y-\mu)$ is said to have the extreme value distribution.

Since $X = -\log(Y - \mu)$ we have by Theorem 2.2.1 that

$Q_X(u) = -\log Q_{Y-\mu}(1 - u)$ and

$$F(y) = \exp(-\exp\{-[(y - \mu')/\sigma']\}), -\infty < y < \infty ,$$

$$Q(u) = \mu' + \sigma'\{-\log[-\log(u)]\}, \ 0 \leq u \leq 1 ,$$

$$f_o(y) = \exp\{-[y + \exp(-y)]\} ,$$

and

$$f_o Q_o(u) = -u \log u$$

where $\mu' = \log \sigma$ and $\sigma' = \gamma$, and σ and γ are the Weibull scale and shape parameters, respectively.

The Cramer Rao Lower Bound (CRLB, see Rao 1973, pp. 324-331) for the variance of unbiased estimators $\hat{\underset{\sim}{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{c})^T$ of $\underset{\sim}{\theta} = (\mu, \sigma, c)^T$ is given by Kübler (1979) as:

$$\mathrm{Var}(\hat{\underset{\sim}{\theta}}) \geq 1/n \ I^{-1}(\underset{\sim}{\theta})$$

where for matrices A and B the notation $A \geq B$ means that $A - B$ is positive semidefinite and $I(\underset{\sim}{\theta}) = (I_{ij}(\underset{\sim}{\theta}))$ is the Fisher information matrix (see Rao 1973, p331) of μ, σ, c and is given by

$$I_{11}(\underset{\sim}{\theta}) = \left(\frac{c-1}{\sigma}\right)^2 \Gamma(h_2) \text{ provided } c > 2 ,$$

$$I_{22}(\underset{\sim}{\theta}) = \left(\frac{c}{\sigma}\right)^2 ,$$

$$I_{33}(\underset{\sim}{\theta}) = H_1 c^{-2} ,$$

$$I_{12}(\underset{\sim}{\theta}) = I_{21}(\underset{\sim}{\theta}) = \frac{c(c-1)}{\sigma^2} \Gamma(h_1) \text{ provided } c > 1 ,$$

$$I_{13}(\underset{\sim}{\theta}) = I_{31}(\underset{\sim}{\theta}) = -\frac{c-1}{c\sigma} \Gamma(h_1) H_2 \text{ provided } c > 1 ,$$

and

$$I_{23}(\underset{\sim}{\theta}) = I_{32}(\underset{\sim}{\theta}) = -\frac{\psi(2)}{\sigma}$$

where

$$H_1 = \psi'(1) + \psi^2(2) \doteq 1.82368066,$$

$$H_2 = \psi(h_1) + 1 ,$$

$$h_j = 1 - j/c , \; j = 1, 2,$$

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \text{ is the Gamma function,}$$

$$\psi(x) = d \log \Gamma(x)/dx = \Gamma'(x)/\Gamma(x) ,$$

and

$$\psi'(x) = d\, \psi(x)/dx \quad .$$

3. IDENTIFICATION OF DISTRIBUTIONAL SHAPE

The identification of the shape of an unknown distribution is the first stage of analysis as we perceive the k sample quantile regression problem. In this section we describe three quantile function approaches to identifying a distributional shape.

In 3.1 we discuss quantile-box plots and present the nonparametric data modeling and goodness-of-fit procedures for one population developed by Parzen (1979a) to determine Q_o. In 3.2 we discuss a parametric approach to estimating the shape parameter γ in the model:

$$Q_i(u) = \mu_i + \sigma_i(Q_o^*(u))^\gamma \ , \ i = 1, \ldots, k \ ,$$

where $Q_o^*(\cdot)$ is a completely specified quantile function. The procedure is a generalization of one proposed by Dubey (1967) for estimating the shape parameter of the Weibull distribution. In 3.3 a procedure for either determining Q_o for k populations or for estimating γ is discussed. The procedure is based on the goodness-of-fit procedures of Parzen.

3.1 Determination of Q_o

In this section we describe the quantile-box plot approach (Parzen 1979a) to represent data and compare k samples of data. We discuss how quantile-box plots can assist one in determining Q_o for the model

$$Q_i(u) = \mu_i + \sigma_i Q_0(u) \quad , \; i = 1, \ldots, k \quad .$$

The nonparametric data modeling techniques and goodness of fit procedures for one population developed by Parzen (1979a) are also discussed.

Quantile-box plots are described by Parzen (1979a) as a "quick and dirty" approach to exploratory data analysis. The technique is a variation of the box and whiskers technique introduced by Tukey (1977). Quantile-box plots assist one in determining the qualitative characteristics of Q(u), e.g. skewness, symmetry, modality, and tail behavior. However the study of quantile-box plots is an imprecise science; much of the interpretative value of the plots, especially for small sample sizes, depends on the predilections of the investigator.

A quantile-box plot of a sample of data consists of a graph of $\tilde{Q}(u)$ (we use $\tilde{Q}_S(u)$) as a function on the unit interval $0 \leq u \leq 1$) on which a series of boxes is superimposed. The boxes have as vertices $(p, \tilde{Q}(p)), (p, \tilde{Q}(1-p)), (1-p, \tilde{Q}(1-p))$, and $(1-p, \tilde{Q}(p))$. One usually chooses p = .25, .125, and .0625 . Within the H box (p = .25) one can draw a horizontal median line through $\tilde{Q}(.5)$. Parzen (1979b,p.243) gives an approximate 95% confidence interval for the median:

$$\tilde{Q}(.5) \pm (2/\sqrt{n})[\tilde{Q}(.75) - \tilde{Q}(.25)] \; .$$

One can use the quantile-box plot technique to classify the distribution of data as normal shaped, skewed right vs. skewed left, or long-tailed vs. short-tailed. One can detect modes as flat spots in $\tilde{Q}(u)$ and the presence of two groups as jumps in $\tilde{Q}(u)$. Intervals of sharp rise outside the D box ($p = .0625$) cause one to suspect the presence of outliers or a long-tailed distribution. Skewness and symmetry can be checked by inspecting the shape of $\tilde{Q}(u)$ within the boxes and also by examining the position of the H box within the E box ($p = .125$) and the E box within the D box.

One can use multiple quantile-box plots to check if k samples of data have homogeneous shapes except for a location-scale shift. Figure 6.C (p. 87) shows the quantile-box plots for four samples of the Hogg (1975) professors' salary data. Comments on the plots are given in Section 6.1.

Parzen's approach to determining Q_o. For a random sample $\{X_1, \ldots, X_n\}$ of a continuous random variable X with cdf F(x) and quantile function Q(u), one hypothesizes a location-scale model

$$H_o: Q(u) = \mu + \sigma Q_o(u) . \qquad (3.1.1)$$

Parzen (1979a) discusses procedures which provide a test of H_o and also yield estimators of the true fQ function when H_o is rejected. The situation of interest is when Q_o is unknown and one would like to test H_o for various specifications of Q_o (e.g. normal vs. logistic vs. Cauchy).

Parzen defines the following quantities:

1) the transformation density, d(u), defined by

$$d(u) = (1/\sigma_o)\, f_o Q_o(u) q(u) \quad , \quad 0 \leq u \leq 1 \ ,$$

where

$$\sigma_o = \int_0^1 f_o Q_o(u) q(u)\, du \ ,$$

$$f_o Q_o(u) = f_o(Q_o(u)),$$

$$q(u) = dQ(u)/du \quad ;$$

2) the transformation distribution, D(u), defined by

$$D(u) = \int_0^u d(t)dt \quad ;$$

3) the complex-valued transformation correlations, $\rho(v)$, $v = 0, \pm 1, \pm 2, \ldots,$ defined by

$$\rho(v) = \int_0^1 \exp(2\pi iuv) d(u) du \quad .$$

One can estimate the above quantities using:

4) $\tilde{d}(u) = (1/\tilde{\sigma}_o) f_o Q_o(u) \tilde{q}(u)$, $0 \leq u \leq 1$, where

$$\tilde{\sigma}_o = \int_0^1 f_o Q_o(u) \tilde{q}(u) du \ ,$$

$\tilde{q}(u)$ is defined by 2.1.2 ;

5) $\tilde{D}(u) = \int_0^u \tilde{d}(t)dt$;

$$6)\quad \tilde{\rho}(v) = \int_0^1 \exp(2\pi iuv)\,\tilde{d}(u)du,\ v = 0, \pm 1, \ldots \ .$$

One can obtain smoothed estimators of the above quantities using antoregressive methods:

$$7)\quad \hat{d}_m(u) = \hat{K}_m \left| \hat{g}_m(\exp(2\pi iu)) \right|^{-2} \ ,$$

where

$$\hat{g}_m(z) = 1 + \hat{\alpha}_m(1)z + \ldots + \hat{\alpha}_m(m)z^m \ ,$$

$\hat{\alpha}_m(1), \ldots, \hat{\alpha}_m(m)$ satisfy the normal equations

$$\tilde{\rho}(-v) + \hat{\alpha}_m(1)\tilde{\rho}(1 - v) + \ldots + \hat{\alpha}_m(m)\tilde{\rho}(m - v) = 0 \ ,$$

$$v = 1, \ldots, m \ ,$$

$$\hat{K}_m = 1 + \hat{\alpha}_m(1)\tilde{\rho}(1) + \ldots + \hat{\alpha}_m(m)\tilde{\rho}(m) \ ,$$

and m is the order of the autoregressive smoothing;

$$8)\quad \hat{D}_m(u) = \int_0^u \hat{d}_m(t)dt \ ;$$

$$9)\quad \hat{fQ}_m(u) = \frac{\left|\hat{g}_m(\exp(2\pi iu))\right|^2 f_o Q_o(u)}{\int_0^1 \left|\hat{g}_m(\exp(2\pi iu))\right|^2 f_o Q_o(u)\tilde{q}(u)du} \ .$$

Parzen proposes minimizing the CAT criterion defined by

$$CAT(m) = 1/n \sum_{j=1}^{m} \hat{K}_j^{-1} - \hat{K}_m^{-1}$$

to determine an "optimal" order $\hat{m}$ of autoregressive smoothing. One can also select an appropriate order of smoothing by visually checking how well $\hat{D}_m(u)$ fits $\tilde{D}(u)$.

The hypothesis H_o: $Q(u) = \mu + \sigma Q_c(u)$ is equivalent to any of the following hypotheses:

1) $d(u) = 1$,

2) $D(u) = u$,

3) $\rho(v) = 0$ for $v \neq 0$.

The following test statistics could be used to test H_o:

1) $\max \tilde{d}(u)$ or $\int_0^1 \log \tilde{d}(u)du$, $0 \leq u \leq 1$

2) $\max |\tilde{D}(u) - u|$, $0 \leq u \leq 1$

3) $|\tilde{\rho}(v)|^2$, $v \neq 0$.

Parzen (1979a) provides references for the properties of these statistics. When CAT selects $\hat{m} = 0$, Parzen regards it as confirmation of H_o.

A useful diagnostic discussed by Parzen (1980) is the p mode or mode percentile. It is defined to be the value of u at which fQ(u) achieves its mode (or maximum value) when fQ(·) is unimodal. When the p mode exceeds .5 the distribution is skewed left and when the p mode is less than .5 the distribution is skewed right.

The function $\widehat{fQ}_m(u)$ is a useful estimator of fQ(u) even when one has sufficient evidence to reject H_o. By examining the interval of u values for which $\hat{D}_m(u)$ (or $\tilde{D}(u)$) is approximately linear in u,

one can detect which parts of the data seem to fit the hypothesized Q_θ function.

The computer package ONESAM (Parzen and White 1979) provides plots of $\tilde{Q}(u)$, $\tilde{q}(u)$, and $\widetilde{fQ}(u)$. By specifying any of several familiar Q_o functions, plots of $\tilde{d}(u)$, $\tilde{D}(u)$, $|\tilde{\rho}(v)|^2$, $\hat{D}_m(u)$, and $\widehat{fQ}_m(u)$ (for several orders m) are produced along with the goodness-of-fit diagnostics discussed above.

3.2 Estimation of the Shape Parameter

Motivated by the fact that Q(u) is of the form

$$Q(u) = \mu + \sigma\ (Q_o^*(u))^\gamma \qquad (3.2.1)$$

for X having the three parameter Weibull distribution (with $Q_o^*(u) = -\log(1-u)$) and the three parameter lognormal distribution (with $Q_o^*(u) = \exp[\Phi^{-1}(u)]$, see Johnson and Kotz 1970, p. 112), we investigate the estimation of the shape parameter in (3.2.1).

We first find an estimator $\tilde{\gamma}$ of γ for the one sample case and then show how to pool estimators $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_k$ obtained from samples from k populations, the ith of which has quantile function

$$Q_i(u) = \mu_i + \sigma_i (Q_o^*(u))^\gamma, \qquad (3.2.2)$$

to produce an estimator of γ.

Theorem 3.2.1: Let $0 < u_1 < u_2 < u_3 < 1$ be values satisfying

$$Q_o^*(u_2) = (Q_o^*(u_1) Q_o^*(u_3))^{\frac{1}{2}}.$$

Then

$$\gamma = \frac{2 \log\{[Q(u_3) - Q(u_2)]/[Q(u_2) - Q(u_1)]\}}{\log [Q_o^*(u_3)/Q_o^*(u_1)]} \quad . \tag{3.2.3}$$

Proof:

For $u_1 < u_2 < u_3$ we have

$$Q(u_j) = \mu + \sigma (Q_o^*(u_j)) \quad , \quad j = 1, 2, 3 \quad .$$

Then

$$\frac{Q(u_3) - Q(u_2)}{Q(u_2) - Q(u_1)} = \frac{(Q_o^*(u_3))^\gamma - (Q_o^*(u_2))^\gamma}{(Q_o^*(u_2))^\gamma - (Q_o^*(u_1))^\gamma} \quad .$$

Since $Q_o^*(u_2) = [Q_o^*(u_1)Q_o^*(u_3)]^{\frac{1}{2}}$, then

$$\frac{Q(u_3) - Q(u_2)}{Q(u_2) - Q(u_1)} = \left[\frac{Q_o^*(u_3)}{Q_o^*(u_1)}\right]^{\frac{\gamma}{2}}$$

and

$$\log\{[Q(u_3) - Q(u_2)]/[Q(u_2) - Q(u_1)]\} = \frac{\gamma}{2} \log[Q_o^*(u_3)/Q_o^*(u_1)] \quad .$$

Hence

$$\gamma = \frac{2 \log\{[Q(u_3) - Q(u_2)]/[Q(u_2) - Q(u_1)]\}}{\log[Q_o^*(u_3)/Q_o^*(u_1)]} \quad .$$

Theorem 3.2.2: Let

$$\gamma = \frac{2\log\{[\hat{Q}(u_3)-\hat{Q}(u_2)]/[\hat{Q}(u_2)-\hat{Q}(u_1)]\}}{\log[Q_o^*(u_3)/Q_o^*(u_1)]} . \quad (3.2.4)$$

Then $\sqrt{n}(\tilde{\gamma} - \gamma) \xrightarrow{d} N(0, V(\gamma))$

where

$$V(\gamma) = \frac{4}{d^2}[\sigma_{11}d_{21}^2 + \sigma_{22}(d_{21} + d_{32})^2 + \sigma_{33}d_{32}^2$$

$$-2\sigma_{12}(d_{21}^2 + d_{21}d_{32}) + 2\sigma_{13}d_{21}d_{32} - 2\sigma_{23}(d_{32}^2 + d_{21}d_{32})], \quad (3.2.5)$$

$$d = \log[Q_o^*(u_3)/Q_o^*(u_1)] ,$$

$$d_{ij} = 1/[(Q_o^*(u_i))^\gamma - (Q_o^*(u_j))^\gamma] ,$$

$$\sigma_{ij} = \frac{\min(u_i, u_j) - u_i u_j}{f_o Q_o^*(u_i) f_o Q_o^*(u_j)} .$$

where $f_o Q_o^*$ is the fQ function corresponding to Q_o^* .

Proof: By Theorem 2.1.4, we have

$$\sqrt{n}(\hat{Q}(u_1) - Q(u_1), \hat{Q}(u_2) - Q(u_2), \hat{Q}(u_3) - Q(u_3))^T \xrightarrow{d} N_3(\underset{\sim}{0}_3, C)$$

where

$$C_{ij} = C_{ji} = u_i(1 - u_j)/(fQ(u_i)fQ(u_j)) , \ 1 \le i \le j \le 3 .$$

Then since $\gamma = g(Q(u_1), Q(u_2), Q(u_3))$ and $\tilde{\gamma} = g(\hat{Q}(u_1), \hat{Q}(u_2), \hat{Q}(u_3))$ where $g(x_1, x_2, x_3)$ is defined by (3.2.3) and (3.2.4), we have (see Rao 1973, p. 387)

$$\sqrt{n}(\tilde{\gamma} - \gamma) \xrightarrow{d} N(0, V(\gamma))$$

where

$$V(\gamma) = \underset{\sim}{t}^T C \underset{\sim}{t} \quad ,$$

$$\underset{\sim}{t} = (\partial g/\partial x_1, \partial g/\partial x_2, \partial g/\partial x_3)^T \quad .$$

Since

$$\partial g/\partial x_1 = 2/[d(Q(u_2) - Q(u_1))] \quad ,$$

$$\partial g/\partial x_2 = -2/d[1/(Q(u_3) - Q(u_2)) + 1/(Q(u_2) - Q(u_1))] \quad ,$$

$$\partial g/\partial x_3 = 2/[d(Q(u_3) - Q(u_2))] \quad ,$$

the theorem follows.

Remarks on Theorem 3.2.2:

1) The estimator $\tilde{\gamma}$ and its asymptotic distribution is independent of μ and σ .

2) Theoretically one can choose optimal values of u_1, u_2, u_3 which minimize the variance of $\tilde{\gamma}$. The values will be a function of γ for a given Q_o^*. Table 3.1 gives optimal values of u_1, u_2, u_3 which minimize $V(\gamma)$ for γ = .05, .1(.1)1 2, 3 when $Q_o^*(u) = -\log(1 - u)$ (i.e. the Weibull distribution). The table also gives the minimum value of $V(\gamma)$ and compares it to the CRLB for unbiased estimators of γ . Figure 3.A plots the optimal values of u_1, u_2, u_3 as a function of γ. See page 32 for further discussion of Table 3.1.

3) Since $V(\gamma)$ is a continuous function of γ, a consistent esti-

mator $\tilde{V}(\gamma)$ of $V(\gamma)$ is obtained by substituting $\hat{\gamma}$ for γ in (3.2.5).

Dubey (1967) gives the formula for an estimator of the shape parameter of the three parameter Weibull distribution when σ and μ are unknown. The estimator of $1/\gamma = c$ is given by

$$(\tilde{1/\gamma}) = \tilde{c} = \frac{\log[-\log(1-u_3)] - \log[-\log(1-u_1)]}{2[\log(\hat{Q}(u_3) - \hat{Q}(u_2)) - \log(\hat{Q}(u_2) - \hat{Q}(u_1))]} ,$$

where

$$u_2 = 1 - \{\exp -[\log(1-u_1)\log(1-u_3)]^{\frac{1}{2}}\} ,$$

which is just the reciprocal of (3.2.4) using $Q_o^*(u) = -\log(1-u)$. Dubey states that the variance of $\tilde{c}$ depends on the true value of c and consequently he does not utilize optimal values of u_1 and u_3 which minimize the variance of $\tilde{c}$.

When one has samples from k populations which satisfy the model (3.2.2), we now show a method to test for homogeneity of shape and to estimate the common value of γ. Let $\hat{Q}_i(u)$ be the empirical quantile function based on a sample of size n_i from population i. To combine estimators $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$ of γ we have

Theorem 3.2.3:

$$\text{Let } H = \sum_{i=1}^{k} \frac{n_i(\hat{\gamma}_i - \hat{\gamma}_p)^2}{V(\gamma)}$$

and

$$\tilde{\gamma}_p = \sum_{i=1}^{k} n_i \tilde{\gamma}_i / n \ , \qquad (3.2.6)$$

where

$$\tilde{\gamma}_i = \frac{2 \log\{[\tilde{Q}_i(u_3) - \tilde{Q}_i(u_2)]/[\tilde{Q}_i(u_2) - \tilde{Q}_i(u_1)]\}}{\log[Q_o^*(u_3)/Q_o^*(u_1)]} \ ,$$

$$n = \sum_{i=1}^{k} n_i \ ,$$

$$Q_o^*(u_2) = [Q_o^*(u_1)\ Q_o^*(u_3)]^{\frac{1}{2}} \ , \ u_1 < u_2 < u_3 \ ,$$

and $V(\gamma)$ is given by (3.2.5) .

If the k populations do in fact have the same shape parameter γ , then

1) $H \xrightarrow{d} \chi^2_{k-1}$,

2) $\sqrt{n}(\tilde{\gamma}_p - \gamma) \xrightarrow{d} N(0, V(\gamma))$

where as $n \to \infty$, the ratio n_i/n approaches a constant.

Proof: This is a direct application of Rao (1973, p. 389).

Remarks on Theorem 3.2.3:

1) The statistic H can be used to test for homogeneity of shape.

2) $\tilde{\gamma}_p$ is a pooled estimator of the common shape parameter of the Q_i's.

3) A consistent estimator of $V(\gamma)$ is obtained by substituting $\tilde{\gamma}_p$ for γ in (3.2.5).

4) Optimal values of u_1, u_2, and u_3 which minimize the variance of $\tilde{\gamma}_p$ will depend on the true value of γ. Table 3.1 can be used to find the optimal values of u_1, u_2, u_3 for a range of values of γ and $Q_o^*(u) = -\log(1 - u)$. Information obtained from quantile-box plots or historical data regarding the distributional shape may help to determine an appropriate set of values of u_1, u_2, u_3.

Remarks on Table 3.1 and Figure 3.A:

Table 3.1 gives the optimal values of u_1, u_2, u_3 for the estimator $\tilde{\gamma}$ assuming Q_o^* for the Weibull distribution, $V(\gamma)$ using these u values, the CRLB for γ when appropriate, and the asymptotic relative efficiency (ARE) of $\tilde{\gamma}$ defined by $ARE(\tilde{\gamma}) = CRLB/V(\gamma)$. Figure 3.A plots the optimal values of u_1, u_2, u_3 as a function of γ. The following trends are evident.

1) For $\gamma > 1.0$ (i.e. the distribution has no mode and is highly skewed right), then $u_3 = 1.0$. The optimal value of u_1 goes from .63 (for γ=3.0) to .01 (for $\gamma = 1.0$).

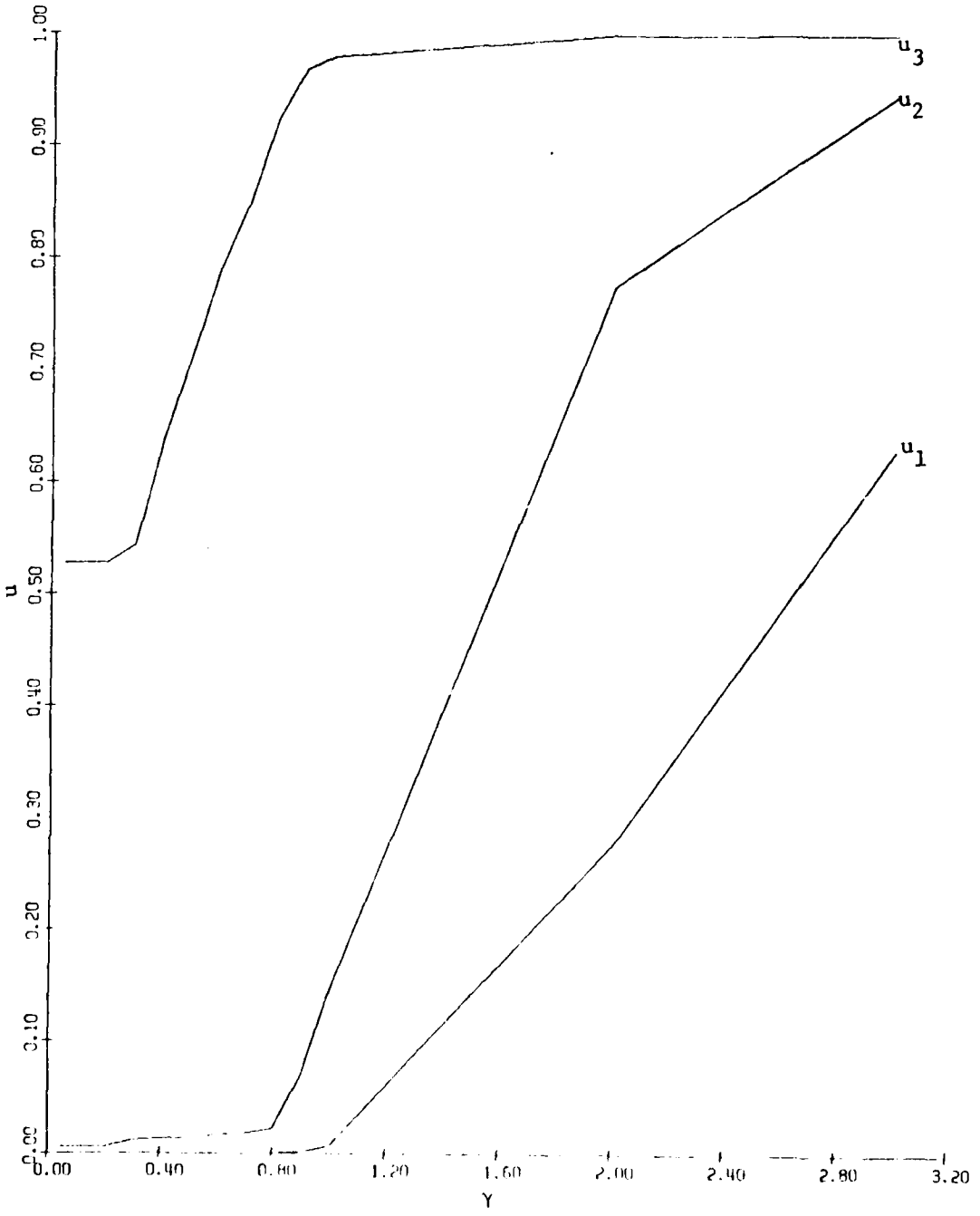
2) For $\gamma = 1.0$ (i.e. the exponential distribution), $u_3 = .98$ and $u_1 = .01$ are optimal.

3) As γ goes from 1.0 to .3 (i.e. the distribution is unimodal and goes from skewed right to almost symmetric),

Table 3.1 Optimal Values of u_1, u_2, u_3 for $\tilde{\gamma}$; Weibull Distribution

γ	u_1	u_2	u_3	$V(\gamma)$	CRLB	$ARE(\tilde{\gamma})$
3	.6283	.9452	.9998	.0341	NA	--
2	.2754	.7750	.9990	.4948	NA	--
1	.0064	.1461	.9795	1.0326	.5483	.531
.9	.0015	.0694	.9690	.8257	.4442	.538
.8	.0002	.0224	.9225	.5799	.3509	.605
.7	.00017	.0177	.8473	.3356	.2687	.8007
.6	.00017	.0162	.7902	.2059	.1974	.9587
.5	.00017	.0145	.7141	.1481	.1371	.9257
.4	.00017	.0131	.6380	.1256	.0877	.6982
.3	.00017	.0115	.5429	.1211	.0494	.4079
.2	.00010	.0057	.5274	.427	.0219	.1943
.1	.0001	.0057	.5274	.4645	.0055	.0118

Figure 3.A Optimal Values of u_1, u_2, u_3 for $\tilde{\gamma}$; Weibull distribution



the optimal value of u_3 goes from .98 (for $\gamma = 1.0$) to .54 (for $\gamma = .3$). The optimal value of u_1 remains close to 0.

4) When $\gamma = .3$ (i.e. the distribution is almost symmetric, normal shaped), $u_1 \doteq .002$ and $u_3 \doteq .54$ are optimal.

5) As γ goes from .3 to .05 (i.e. the distribution goes from almost symmetric to skewed left), the optimal value of u_3 goes from .54 (for $\gamma = .3$) to .52 (for $\gamma = .05$) and the optimal value of u_1 remains close to 0.

6) The ARE increases from .53 for $\gamma = 1.0$ to .96 for $\gamma = .6$. The ARE is .41 when $\gamma = .3$ and decreases rapidly to .01 for $\gamma = .05$.

7) The CRLB for γ is inappropriate for $\gamma > 1.0$ since the Fisher information measure for $\gamma > 1.0$ does not exist. One might wish to compare $V(\gamma)$ to the asymptotic variance of the maximum likelihood estimator of γ based on a censored sample (see Harter and Moore 1967).

Thus a strategy emerges. If one assumes the data to have a Weibull distribution with an unknown shape parameter, one can select almost optimal values of u_1 and u_3 (and consequently u_2) according to the shape suggested by quantile-box plots or other graphical techniques. If the shape is "super exponential" (i.e. very skewed right and no mode), then select $u_3 \doteq 1.0$ and u_1 in the range (.01, .63) (a longer tail implies a larger value of u_1). If the data seem to be exponential (i.e. skewed right and no mode), choose $u_3 \doteq .97$ and

$u_1 \doteq .0064$. If the data are unimodal and skewed right, values of u_3 in the range (.6, .95) (a longer tail implies a larger value of u_3) and $u_1 \doteq .0002$ will be almost optimal. If the data seem almost symmetric, select $u_3 \doteq .55$ and $u_1 \doteq .0002$. If the data are unimodal and skewed left, values of $u_3 \doteq .54$ and $u_1 \doteq .0002$ will be almost optimal.

The estimation of γ can also be done iteratively. An estimate $\tilde{\gamma}$ based on one set of (u_1, u_2, u_3) values may suggest better values of (u_1, u_2, u_3).

In Section 6.2 we illustrate the use of the estimator $\tilde{\gamma}_p$ of γ for ten samples of data representing the tolerance of green sunfish to thermal pollution.

3.3 A Goodness-of-Fit Approach for Determining Distributional Shape

In this section we describe how to apply the one population goodness-of-fit (GOF) procedures of Section 3.1 to the estimation of the common value of γ in the k population model

$$Q_i(u) = \mu_i + \sigma_i \, [Q_o^*(u)]^{\gamma} \ , \ i = 1, \ldots, k \ ,$$

where $Q_o^*(\cdot)$ is assumed known and γ is an unknown shape parameter, and also to the identification of Q_o in the model

$$Q_i(u) = \mu_i + \sigma_i \, Q_o(u) \ , \ i = 1, \ldots, k \ .$$

Estimation of γ using GOF. The proposed GOF procedure consists of three parts:

1) Form a grid $\{\gamma_{o1}, \ldots, \gamma_{om}\}$ of potential values for γ.

2) For each value of γ_o in the grid, form estimates $\hat{\mu}_i(\gamma_o)$ and $\hat{\sigma}_i(\gamma_o)$ of μ_i and σ_i using linear combinations of order statistics (see Section 4.2 below). For the ith sample form a transformed sample $\tilde{Q}_{i,\gamma_o}(u)$ defined by

$$\tilde{Q}_{i,\gamma_o}(u_j) = \left[\frac{\tilde{Q}_i(u_j) - \hat{\mu}_i(\gamma_o)}{\hat{\sigma}_i(\gamma_o)}\right]^{1/\gamma_o} ,$$

$$u_j = (j - .5)/n_i \ , \ j = 1, \ldots, n_i, \ i = 1, \ldots, k \ .$$

Next, pool the k transformed samples.

3) The hypothesis

$$H_o: \ Q_i(u) = \mu_i + \sigma_i(Q_o^*(u))^{\gamma_o}$$

can be written as

$$H_o: \left[\frac{Q_i(u) - \mu_i}{\sigma_i}\right]^{1/\gamma_o} = Q_o^*(u) \ .$$

Under this hypothesis for a specified value of γ_o, we can consider the pooled transformed sample as a random sample of size $n = \sum_{i=1}^{k} n_i$ from a population with quantile function $Q_o^*(u)$. Select the best value of γ_o to be the value that gives the best agreement of the pooled transformed sample and $Q_o^*(u)$ according to the GOF criteria of Section 3.1.

Considerations in determining the grid $\{\gamma_{o1}, \ldots, \gamma_{om}\}$ are:

1) Knowledge of the qualitative properties of $Q_i(u)$ (e.g. symmetry, right skewness, or left skewness, and tail behavior) obtained from quantile-box plots can help one select a narrow grid of γ_o values.
2) Computationally the procedure is expensive.
3) If one specifies $Q_o^*(u) = -\log(1 - u)$ (i.e. $Q_i(u)$ is in the Weibull family), tables of optimal spacings and coefficients for the estimation of μ_i and σ_i using linear combination of order statistics are available for limited values of γ. Programs to compute the optimal spacings and coefficients for a wide range of γ values should be made available.

Identification of Q_o using GOF. The above procedure is also appropriate for identifying Q_o. The steps are:

1) Specify a family of possible Q_o functions.
2) For each Q_o function form estimates $\hat{\mu}_i(Q_o)$ and $\hat{\sigma}_i(Q_o)$ of μ_i and σ_i using linear combinations of order statistics. For the ith sample form a transformed sample defined by

$$\tilde{Q}_{i,Q_o}(u) = \frac{Q_i(u) - \hat{\mu}_i(Q_o)}{\hat{\sigma}_i(Q_o)} \quad , i = 1, \ldots k$$

and pool the k transformed samples.
3) The hypothesis

$$H_o: \; Q_i(u) = \mu_i + \sigma_i Q_o(u) \quad , i = 1, \ldots, k$$

is equivalent to

$$H_o: \frac{Q_i(u) - \mu_i}{\sigma_i} = Q_o(u) \quad .$$

Thus under this hypothesis for a specified Q_o, we can consider the pooled transformed sample as a random sample of size $n = \sum_{i=1}^{k} n_i$ from a population with quantile function $Q_o(u)$. Select the best specification of Q_o as the one that gives the best agreement of the pooled transformed sample and $Q_o(u)$ according to the criteria of Section 3.1.

Examination of how the misspecification of γ_o affects the estimators $\hat{\mu}$ and $\hat{\sigma}$ for the one population case is examined analytically in Section 4.3. In Sections 6.1 and 6.2 we illustrate the techniques of identifying Q_o using the professors' salary data of Hogg (1975) and the green sunfish data of Matis and Wehrly (1979).

4. ESTIMATION OF LOCATION AND SCALE PARAMETERS

The estimation of location and scale parameters is the second stage in the k-sample quantile regression procedure as we perceive it. Existing techniques for the estimation of location and scale parameters in the one population case are also appropriate for the estimation of location and scale parameters when there are k populations satisfying

$$Q_i(u) = \mu_i + \sigma_i Q_o(u) \ , \ i = 1, \ldots, k \ ,$$

where μ_i and σ_i are the location and scale parameters respectively of the ith population.

The location and scale parameter model for one population can be written

$$F(x) = F_o\left(\frac{x - \mu}{\sigma}\right)$$

or

$$Q(u) = \mu + \sigma Q_o(u)$$

where F_o and Q_o are completely specified and μ and σ are unknown location and scale parameters, respectively. One would like estimators of μ and σ which are statistically efficient and relatively simple to compute.

Estimators $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ based on linear combinations of order statistics (LCOS) are defined by

$$\hat{\mu} = \sum_{j=1}^{r} a_j \tilde{Q}(u_j)$$

$$\hat{\sigma} = \sum_{j=1}^{r} b_j \tilde{Q}(u_j)$$

where $\tilde{Q}(u)$ is the empirical quantile function, r is the number of values of $\tilde{Q}(u)$ (or the number of order statistics) used, and a_j, b_j, $j = 1, \ldots, r$, are specified constants. Two approaches to the choice of r, the a_j's and b_j's, and u_j's are discussed in the next two sections. The first approach (Section 4.1) is due to Ogawa (1951) and Hassanein (1971, 1972) and the second (Section 4.2) is due to Eubank (1979). Section 4.3 investigates the estimation of μ and σ for the Weibull distribution when the shape parameter γ is misspecified.

4.1 Optimal Linear Combinations of Order Statistics

In this section we present the general work of Ogawa (1951) and the work of Hassanein (1971, 1972) dealing with the selection of optimal linear combinations of order statistics for the simultaneous estimation of μ and σ.

Using the model

$$Q(u) = \mu + \sigma Q_o(u)$$

and recalling Theorem 2.1.4 regarding the asymptotic distribution of $\tilde{Q}(u)$, asymptotically $\tilde{Q}(u_1), \ldots, \tilde{Q}(u_r)$ satisfy the conditions required for the application of the Gauss-Markov Theorem. Thus generalized least squares may be used to obtain asymptotically best linear unbiased estimators (ABLUE's) of μ and/or σ. Ogawa (1951, see

Eubank 1979) gives general formulae for the estimators and their asymptotic relative efficiencies(ARE's) when compared to the CRLB.

Let $u_0 = 0$, $u_{r+1} = 1$ and $f_oQ_o(u_0) = f_oQ_o(u_{r+1}) = 0$.

Define

$$K_{11} = \sum_{j=1}^{r+1} \frac{[f_oQ_o(u_j) - f_oQ_o(u_{j-1})]^2}{u_j - u_{j-1}} , \qquad (4.1.1)$$

$$K_{22} = \sum_{j=1}^{r+1} \frac{[Q_o(u_j)f_oQ_o(u_j) - Q_o(u_{j-1})f_oQ_o(u_{j-1})]^2}{u_j - u_{j-1}} , \qquad (4.1.2)$$

$$K_{12} = \sum_{j=1}^{r+1} \frac{[f_oQ_o(u_j)-f_oQ_o(u_{j-1})][Q_o(u_j)f_oQ_o(u_j)-Q_o(u_{j-1})f_oQ_o(u_{j-1})]}{u_j - u_{j-1}} , \qquad (4.1.3)$$

$$\Delta = K_{11}K_{22} - K_{12}^2 ,$$

$$K_{01} = \sum_{j=1}^{r+1} \frac{[f_oQ_o(u_j)-f_oQ_o(u_{j-1})][f_oQ_o(u_j)\tilde{Q}(u_j)-f_oQ_o(u_{j-1})\tilde{Q}(u_{j-1})]}{u_j - u_{j-1}} ,$$

$$K_{02} = \sum_{j=1}^{r+1} \frac{1}{u_j - u_{j-1}} \left\{ \begin{matrix} [Q_o(u_j)f_oQ_o(u_j)-Q_o(u_{j-1})f_oQ_o(u_{j-1})] \\ [f_oQ_o(u_j)\tilde{Q}(u_j)-f_oQ_o(u_{j-1})\tilde{Q}(u_{j-1})] \end{matrix} \right\}$$

Then ABLUE's for the simultaneous estimation of μ and σ are given by

$$\hat{\mu} = (K_{22}K_{01} - K_{12}K_{02})/\Delta , \qquad (4.1.4)$$

$$\hat{\sigma} = (K_{11} K_{02} - K_{12} K_{01})/ \Delta . \qquad (4.1.5)$$

Notice that these are just the generalized least squares estimators of μ and σ given by

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = (X^T C^{-1} X)^{-1} X^T C^{-1} \underset{\sim}{Y} ,$$

where

$$X^T C^{-1} X = \begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix}$$

$$X^T C^{-1} \underset{\sim}{Y} = \begin{pmatrix} K_{01} \\ K_{02} \end{pmatrix} .$$

where C is given by (2.1.3).

For the simultaneous estimation of μ and σ Ogawa (1951) defines the ARE by

$$ARE(\hat{\mu}, \hat{\sigma}) = \frac{|I^{-1}(\underset{\sim}{\theta})|}{Var(\hat{\mu})Var(\hat{\sigma}) - Cov^2(\hat{\mu}, \hat{\sigma})} ,$$

where

$$\underset{\sim}{\theta} = (\mu, \sigma)^T$$

$$I(\underset{\sim}{\theta}) = \begin{bmatrix} E\left[\left(\frac{f'(X)}{f(X)}\right)^2\right] & E\left[X\left(\frac{f'(X)}{f(X)}\right)^2\right] \\ E\left[X\left(\frac{f'(X)}{f(X)}\right)^2\right] & E\left[\left(\frac{X f'(X)}{f(X)}\right)^2\right] - 1 \end{bmatrix}$$

$$= \begin{bmatrix} \langle f_o Q_o(u),\ f_o Q_o(u) \rangle & \langle f_o Q_o(u),\ Q_o(u) f_o Q_o(u) \rangle \\ \langle f_o Q_o(u),\ Q_o(u) f_o Q_o(u) \rangle & \langle Q_o(u) f_o Q(u),\ Q_o(u) f_o Q_o(u) \rangle \end{bmatrix} \tag{4.1.6}$$

and

$$\langle f(u),\ g(u) \rangle = \int_0^1 f'(u)\ g'(u)\ du \quad .$$

Examination of the equations for the estimators and their ARE's reveals that the equations are functions of the spacings $u_1, \ldots, u_r$. Thus the problem reduces to finding a set of optimal spacings which maximize the ARE of the estimators. For certain distributions the expressions for the ARE's are quite complicated and numerical methods have to be used to find optimal or near optimal spacings. For a given distribution the results are usually expressed as tables of optimal spacings $u_1, \ldots, u_r$ and the corresponding coefficients $a_1, \ldots, a_r, b_1, \ldots, b_r$ for the ABLUE's for various values of r.

Hassanein (1971) uses this procedure to find optimal spacings and coefficients for the simultaneous estimation of μ and σ for the Weibull distribution. The tables he provides are a function of the shape parameter, $c = 1/\gamma$, and he provides spacings and coefficients for r = 2, 4, 6 order statistics. The values of c he considers are c = 3(1)10(5)20 . Subroutine QTOLSW uses Hassanein's tables for r = 6 values to compute estimates of μ and σ for any of the above specified values of c.

Hassanein (1972) considers the problem of selecting optimal spacings and coefficients for the simultaneous estimation of the location and scale parameters of the extreme value distribution.

He provides optimal spacings for r = 1(1)10 order statistics. Let us recall the property that if X has a Weibull distribution with $\mu = 0$, then Y = log X has an extreme value distribution with the location parameter $\mu' = \log \sigma$ and $\sigma' = \gamma$ where σ and γ are the scale and shape parameters of the Weibull distribution. Thus one can use the optimal spacings and coefficients for the extreme value distribution to estimate the scale and shape parameters of the Weibull distribution as long as the location parameter μ is known.

There is an extensive literature on the use of linear combinations of order statistics to estimate location and scale parameters for many common distributions. The approach adopted by Ogawa, Hassanein, and others centers on maximizing the ARE of the estimators. In the next section we present another approach to the selection of a set of spacings and coefficients for optimal location and scale parameter estimation.

4.2 Asymptotically Optimal Linear Combinations of Order Statistics

In this section we discuss the approach taken by Eubank (1979) for the selection of asymptotically optimal LCOS for the simultaneous estimation of μ and σ. Eubank formulates the problem within the framework of continuous parameter time series regression. Using Theorem 2.1.3 and the model

$$Q(u) = \mu + \sigma Q_0(u) ,$$

we have

$$\sqrt{n}/\sigma \; f_o Q_o(u)[\tilde{Q}(u) - \mu - \sigma \, Q_o(u)] \xrightarrow{d} B(u) \qquad (4.2.1)$$

where $\{B(u), u \in [0, 1]\}$ is a Brownian bridge process.

Then we can write a regression model

$$f_o Q_o(u)\tilde{Q}(u) = \mu f_o Q_o(u) + \sigma Q_o(u) f_o Q_o(u) + \sigma_B B(u) \qquad (4.2.2)$$

where $\sigma_B = \sigma/\sqrt{n}$ is estimated as a free parameter and is not constrained to be related to σ. Eubank restates the problem of selecting a set of spacings for the estimation of μ and σ as that of selecting an optimal design for a Brownian bridge process.

Definition 4.2.1: An r point design for a Brownian bridge process, and consequently for $\{f_o Q_o(u)\tilde{Q}(u), 0 \leq u \leq 1\}$, is an r-tuple $\{u_1, \ldots, u_r\}$ with $0 < u_1 < \ldots < u_r < 1$. Denote by D_r the set of all such r point designs.

Definition 4.2.2: For $T \in D_r$, let $\hat{\underset{\sim}{\theta}}_T$ denote the best linear unbiased estimator (BLUE) of $\underset{\sim}{\theta} = (\mu, \sigma)$ based on observations taken according to T. Let $\hat{\underset{\sim}{\theta}}$ denote the estimator of $\underset{\sim}{\theta}$ obtained using observations over all of [0, 1].

Definition 4.2.3: A design sequence $\{T_r\}_{r=1}^{\infty}$, $T_r \in D_r$, is asymptotically optimal for estimating $\underset{\sim}{\theta}$ if

$$\lim_{r\to\infty} \frac{|\mathrm{Var}^{-1}(\hat{\underset{\sim}{\theta}}_{T_r})| - |\mathrm{Var}^{-1}(\hat{\underset{\sim}{\theta}})|}{\inf_{T\in D_r} |\mathrm{Var}^{-1}(\hat{\underset{\sim}{\theta}}_T)| - |\mathrm{Var}^{-1}(\hat{\underset{\sim}{\theta}})|} = 1 \quad .$$

Theorem 4.2.1 (Eubank 1979): Suppose $f_oQ_o(u)$ and $Q_o(u)f_oQ_o(u)$ have the representations:

$$f_oQ_o(u) = -\int_0^1 (f_oQ_o(t))'' \; K_B(u, t)dt \quad ,$$

$$Q_o(u)f_oQ_o(u) = -\int_0^1 (Q_o(t)f_oQ_o(t))'' \; K_B(u, t)dt$$

where

$$K_B(u, t) = \min(u, t) - u\,t$$

and $(g(t))''$ denotes $d^2g(t)/d$

$$\underset{\sim}{\psi}(u) = -((f_oQ_o(u))'' \; , (Q_o(\mu)f_oQ_o(u))'')^T \quad .$$

Then the density

$$h(u) = \frac{[\underset{\sim}{\psi}(u)^T \, I^{-1}(\underset{\sim}{\theta}) \, \underset{\sim}{\psi}(u)]^{\frac{1}{3}}}{\int^1 [\underset{\sim}{\psi}(u)^T \, I^{-1}(\underset{\sim}{0}) \, \underset{\sim}{\psi}(u)]^{\frac{1}{3}} du}$$

where $\underset{\sim}{0} = (\mu, \sigma)$ and $I(\underset{\sim}{0})$ is defined by (4.1.6) generates asymptotically optimal designs for the simultaneous estimation of μ and σ.

(see remark 2 on p. 43)

Remarks on Theorem 4.2.1:

1) The asymptotic optimality of the design means that as r,

the number of spacings, grows large, then the spacings generated by Theorem 4.2.1 lead to estimators with approximately the same efficiency as estimators based on the optimal set of r spacings.

2) Optimal designs are those that minimize the generalized variance $|\text{Var}\ (\hat{\mu}_T\ ,\ \hat{\sigma}_T)|$.

3) The density h generates the asymptotically optimal design sequence $\{T_r\}_{r=1}^{\infty}$ where

$$T_r = \{H^{-1}\ (\tfrac{1}{r+1})\ ,\ H^{-1}\ (\tfrac{2}{r+1})\ ,\ \ldots,\ H^{-1}\ (\tfrac{r}{r+1})\}$$

and

$$H(u) = \int_0^u h(t)\ dt \quad .$$

Eubank (1979) supplies general formulae for the coefficients a_j and b_j for the estimation of μ and σ using the asymptotically optimal spacings to yield estimators

$$\hat{\mu} = \sum_{j=1}^{r} a_j\ \tilde{Q}(H^{-1}(\tfrac{j}{r+1}))$$

$$\hat{\sigma} = \sum_{j=1}^{r} b_j\ \tilde{Q}(H^{-1}(\tfrac{j}{r+1}))$$

where

$$a_j = [K_{22}(h)W_{\mu}(j,h) - K_{12}(h)W_{\sigma}(j,\ h)]\ /\ \Delta(h)$$

$$b_j = [K_{11}(h)W_{\sigma}(j,h) - K_{12}(h)W_{\mu}(j,h)]\ /\ \Delta(h)\ ,$$

$K_{ij}(h)$ is the same as K_{ij} given by (4.1.1), (4.1.2) and (4.1.3) with u_j replaced by $H^{-1}(j/(r+1))$,

$$\Delta(h) = K_{11}(h)K_{22}(h) - [K_{12}(h)]^2 ,$$

$$W_{\mu}(j,h) = \frac{f_o Q_o(H^{-1}(\frac{j}{r+1}))}{K_{11}(h)} \left[\frac{f_o Q_o(H^{-1}(\frac{j}{r+1})) - f_o Q_o(H^{-1}(\frac{j-1}{r+1}))}{H^{-1}(\frac{j}{r+1}) - H^{-1}(\frac{j-1}{r+1})} - \frac{f_o Q_o(H^{-1}(\frac{j+1}{r+1})) - f_o Q_o(H^{-1}(\frac{j}{r+1}))}{H^{-1}(\frac{j+1}{r+1}) - H^{-1}(\frac{j}{r+1})} \right]$$

and

$$W_{\sigma}(j,h) = \frac{f_o Q_o(H^{-1}(\frac{j}{r+1}))}{K_{22}(h)} \times$$

$$\left[\frac{f_o Q_o(H^{-1}(\frac{j}{r+1}))Q_o(H^{-1}(\frac{j}{r+1})) - f_o Q_o(H^{-1}(\frac{j-1}{r+1}))Q_o(H^{-1}(\frac{j-1}{r+1}))}{H^{-1}(\frac{j}{r+1}) - H^{-1}(\frac{j-1}{r+1})} - \frac{f_o Q_o(H^{-1}(\frac{j+1}{r+1}))Q_o(H^{-1}(\frac{j+1}{r+1})) - f_o Q_o(H^{-1}(\frac{j}{r+1}))Q_o(H^{-1}(\frac{j}{r+1}))}{H^{-1}(\frac{j+1}{r+1}) - H^{-1}(\frac{j}{r+1})} \right]$$

While the approach is direct and once H(u) is computed asymptotically optimal spacings can easily be found, many distributions do not satisfy the required representation for $f_o Q_o(u)$ and $Q_o(u) f_o Q_o(u)$. (see Eubank 1979, p. 116)

Eubank (1979) gives tables of asymptotically optimal spacings and coefficients for the simultaneous estimation of μ and σ based on $r = 2, 7, 9$ order statistics for the normal and logistic distributions. It has been suggested (Eubank, personal communication 1980) that

asymptotically optimal spacings for the simultaneous estimation of μ and σ can be generated for the Weibull distribution for certain values of the shape parameter γ.

4.3 Estimation of μ and σ when γ is Misspecified

In this section we examine analytically how the misspecification of γ affects estimators of μ and σ based on LCOS for the Weibull distribution. Using estimators $\hat{\mu}(\gamma_o)$ and $\hat{\sigma}(\gamma_o)$ based on a range of specified values, γ_o, of the true value, γ, of the shape parameter, we have:

$$\text{Bias}(\hat{\mu}(\gamma_o)) = E(\hat{\mu}(\gamma_o)) - \mu$$

$$= \sum_{j=1}^{r} a_j(-\log(1-u_j))^{\gamma},$$

$$\text{Bias}(\hat{\sigma}(\gamma_o)) = E(\hat{\sigma}(\gamma_o)) - \sigma$$

$$= \sum_{j=1}^{r} b_j(-\log(1-u_j))^{\gamma} - 1,$$

$$\text{Var}(\hat{\mu}(\gamma_o)) = \sum_j \sum_i a_i a_j \text{Cov}(\tilde{Q}(u_i), \tilde{Q}(u_j))$$

$$= \frac{\gamma^2}{n} \sum_j \sum_i \frac{a_i a_j [\min(u_i,u_j) - u_i u_j]}{(1-u_i)(1-u_j)[\log(1-u_i)\log(1-u_j)]^{1-\gamma}},$$

$$\text{Var}(\hat{\sigma}(\gamma_o)) = \sum_j \sum_i b_i b_j \text{Cov}(\tilde{Q}(u_i), \tilde{Q}(u_j))$$

$$= \frac{\gamma^2}{n} \sum_j \sum_i \frac{b_i b_j [\min(u_i,u_j) - u_i u_j]}{(1-u_i)(1-u_j)[\log(1-u_i)\log(1-u_j)]^{1-\gamma}},$$

$$MSE(\hat{\mu}(\gamma_o)) = E(\hat{\mu}(\gamma_o) - \mu)^2$$

$$= [Bias(\hat{\mu}(\gamma_o))]^2 + Var(\hat{\mu}(\gamma_o)) \quad ,$$

$$MSE(\hat{\sigma}(\gamma_o)) = E(\hat{\sigma}(\gamma_o) - \sigma)^2$$

$$= [Bias(\hat{\sigma}(\gamma_o))]^2 + Var(\hat{\sigma}(\gamma_o)) \quad .$$

We calculate the values of each of the above properties of $\hat{\mu}(\gamma_o)$ and $\hat{\sigma}(\gamma_o)$ using the values of γ and γ_o: .333, .25, .20, .167, .143, .125, .111, .10, .067 and .05 . The coefficients $\{a_j\}$ and $\{b_j\}$ and optimal values $\{u_j\}$ are obtained from tables given by Hassanein (1971) using r = 6 order statistics and the specified γ_o value. The MSE (mean squared error) is computed for samples of size n = 20 and 50.

The results are summarized in Tables 4.1, 4.2, 4.3, and 4.4. The first entry in each cell of the various tables is for $\hat{\mu}(\gamma_o)$ and the second entry is for $\hat{\sigma}(\gamma_o)$. Figures 4.A, 4.B, 4.C and 4.D represent plots of the properties of the estimators vs. the specified value of γ_o. Each curve on the plots represents a distinct value of γ as indicated in the key. Plotting the curves for all the values of γ on the same set of axes facilitates comparison of the properties for different misspecifications.

Remarks on Table 4.1 and Figure 4.A:

Table 4.1 and Figure 4.A present the results of a bias study of the estimators $\hat{\mu}(\gamma_o)$ and $\hat{\sigma}(\gamma_o)$. The following remarks can be made:

Table 4.1 Bias of $\hat{\mu}(\gamma_o)$, $\hat{\sigma}(\gamma_o)$

	γ_o									
γ	.333	.250	.200	.167	.143	.125	.111	.100	.067	.050
.333	.000	-.239	-.519	-.814	-1.117	-1.424	-1.734	-2.046	-3.618	-5.201
	-.001	.264	.557	.860	1.168	1.479	1.792	2.105	3.684	5.270
.250	.163	.000	-.204	-.422	-.648	-.877	-1.109	-1.342	-2.522	-3.710
	-.185	.000	.215	.440	.670	.902	1.136	1.372	2.556	3.747
.200	.279	.159	.000	-.173	-.353	-.536	-.722	-.909	-1.855	-2.808
	-.308	-.169	.000	.179	.363	.548	.736	.925	1.875	2.830
.167	.367	.274	.144	.000	-.149	-.302	-.457	-.613	-1.404	-2.201
	-.398	-.288	-.149	.000	.153	.308	.465	.662	1.416	2.215
.143	.435	.360	.250	.127	-.001	-.132	-.269	-.409	-1.079	-1.765
	-.466	-.376	-.257	-.130	.001	.134	.269	.404	1.088	1.775
.125	.490	.429	.334	.227	.115	.000	-.116	-.234	-.830	-1.431
	-.520	-.445	-.342	-.231	-.117	.000	.118	.236	.835	1.438
.111	.535	.484	.400	.305	.206	.104	.000	-.105	-.635	-1.171
	-.565	-.499	-.409	-.311	-.209	-.105	.000	-.106	.639	1.176
.100	.573	.529	.454	.369	.280	.188	.094	.000	-.478	-.961
	-.601	-.544	-.463	-.375	-.283	-.190	-.095	.000	.481	.965
.067	.697	.672	.624	.567	.508	.446	.384	.321	.000	-.324
	-.719	-.685	-.632	-.573	-.512	-.450	-.386	-.323	.000	.325
.050	.765	.749	.713	.671	.626	.580	.533	.485	.244	.000
	-.783	-.759	-.720	-.676	-.630	-.583	-.536	-.488	-.245	.000

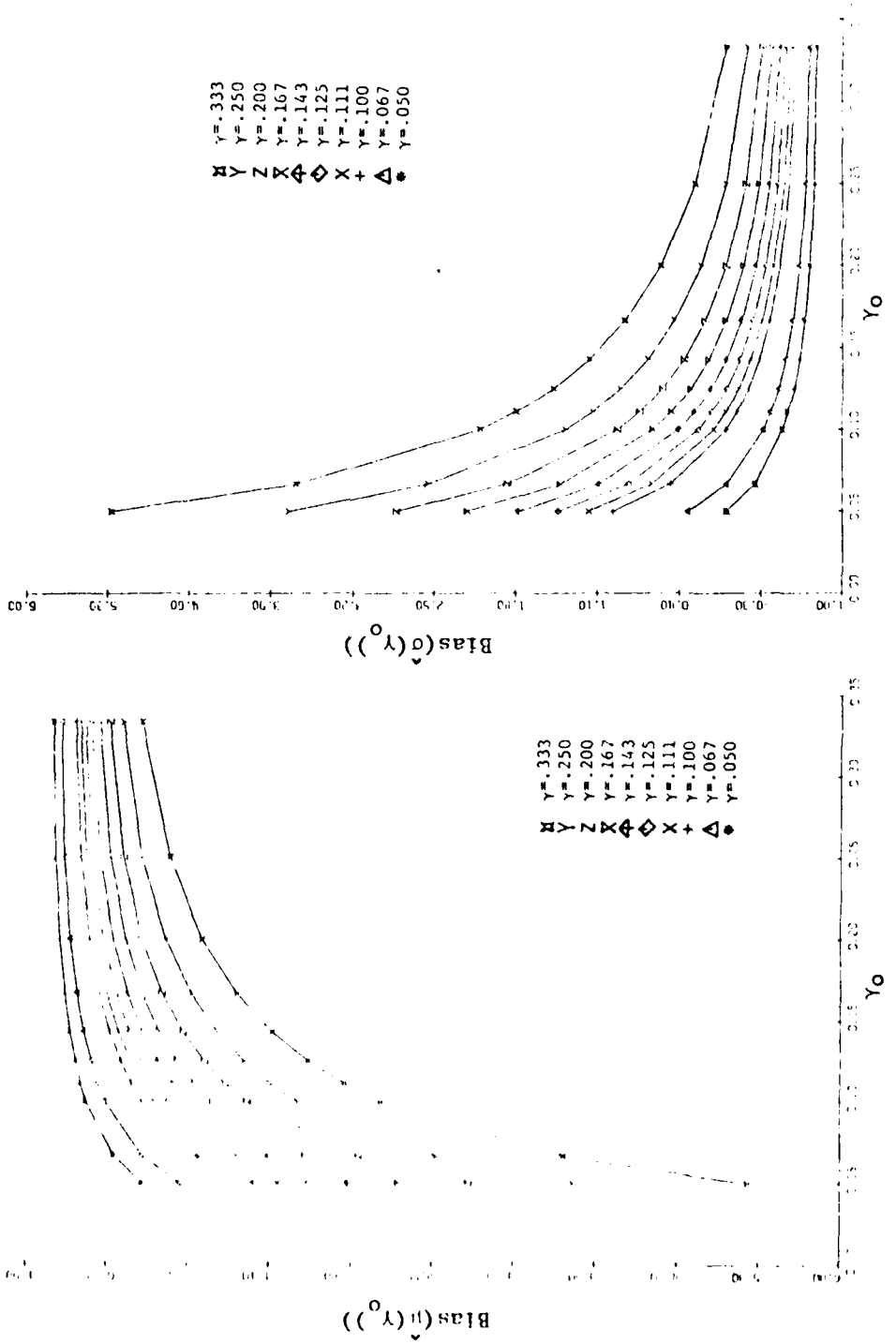


Figure 4.A Bias of $\hat{\mu}(\gamma_o)$, $\hat{\sigma}(\gamma_o)$ as Function of γ

1) Bias $(\hat{\mu}(\gamma_o))$ and Bias $(\hat{\sigma}(\gamma_o))$ for $\gamma = .333$ get large in magnitude as γ_o gets small. A consequence of this is that if the data is almost symmetric and one specifies the data to be skewed left (γ_o small), then $\hat{\mu}(\gamma_o)$ may seriously underestimate μ and $\hat{\sigma}(\gamma_o)$ may seriously overestimate σ.
2) For data that is skewed left, the risk of a seriously biased estimator when γ_o is misspecified is not as great as in (1). When $\gamma = .05$ and one specifies $\gamma_o = .333$, $\text{Bias}(\hat{\mu}(.333)) = .8$ and $\text{Bias}(\hat{\sigma}(.333)) = -.8$.
3) $\text{Bias}(\hat{\mu}(\gamma_o)) = -\text{Bias}(\hat{\sigma}(\gamma_o))$ for all values of γ.
4) One might wish to approximate the bias curves as a function of γ and γ_o. While this would be useful in general, examination of the general formulae for $\hat{\mu}(\gamma_o)$ and $\hat{\sigma}(\gamma_o)$ given by 4.1.4 and 4.1.5 do not offer much promise of this.

Remarks on Table 4.2 and Figure 4.B

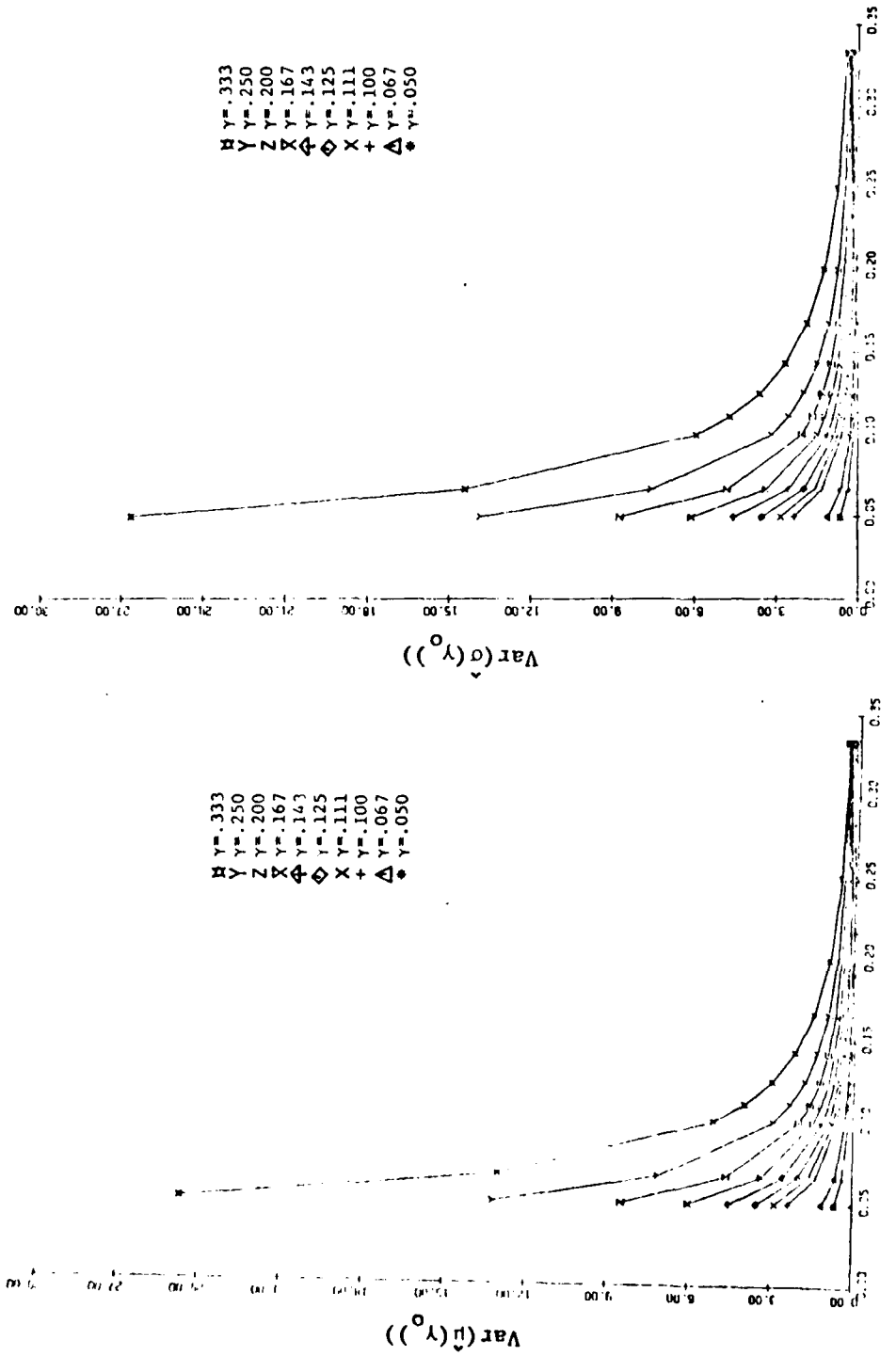
Table 4.2 and Figure 4.B present the results of a study of $\text{Var}(\hat{\mu}(\gamma_o))$ and $\text{Var}(\hat{\sigma}(\gamma_o))$. The following remarks can be made:

1) It should be noted that the table and figure give $n\text{Var}(\hat{\mu}(\gamma_o))$ and $n\text{Var}(\hat{\sigma}(\gamma_o))$ since the variance of each estimator is a function of the sample size.
2) In general $\text{Var}(\hat{\mu}(\gamma_o))$ and $\text{Var}(\hat{\sigma}(\gamma_o))$ remain fairly constant for γ_o equal to .25 or .333 regardless of the true value γ.
3) $\text{Var}(\hat{\mu}(\gamma_o)) = \text{Var}(\hat{\sigma}(\gamma_o))$.

Table 4.2 Variance of $\hat{\mu}(\gamma_o)$, $\hat{\sigma}(\gamma_o)$

	γ_o									
γ	.333	.250	.200	.167	.143	.125	.111	.100	.067	.050
.333	.358	.583	.968	1.502	2.184	3.014	3.994	5.123	13.024	24.688
	.428	.749	1.237	1.877	2.667	3.607	4.696	5.936	14.391	26.612
.250	.401	.472	.683	.930	1.354	1.806	2.335	2.940	7.112	13.192
	.360	.482	.733	1.068	1.482	1.974	2.542	3.186	7.555	13.834
.200	.421	.400	.530	.723	.967	1.260	1.601	1.990	4.646	8.482
	.343	.366	.517	.728	.989	1.300	1.658	2.064	4.804	8.724
.167	.421	.343	.428	.564	.738	.947	1.189	1.465	3.332	6.013
	.331	.297	.396	.542	.724	.941	1.192	1.475	3.382	6.102
.143	.407	.297	.355	.456	.587	.745	.927	1.135	2.534	4.533
	.316	.249	.317	.425	.561	.723	.909	1.121	2.538	4.556
.125	.386	.258	.298	.377	.479	.602	.744	.906	1.995	3.544
	.298	.213	.261	.343	.448	.574	.719	.882	1.979	3.537
.111	.362	.226	.255	.317	.399	.498	.613	.743	1.618	2.859
	.278	.184	.220	.285	.369	.469	.585	.716	1.594	2.838
.100	.337	.200	.220	.270	.338	.420	.514	.622	1.341	2.360
	.259	.161	.188	.240	.309	.391	.486	.594	1.313	2.332
.067	.231	.116	.120	.142	.173	.212	.257	.307	.645	1.120
	.173	.092	.099	.123	.154	.193	.237	.287	.622	1.094
.050	163	.075	.075	.087	.105	.128	.154	.183	.380	.655
	.126	.059	.062	.075	.093	.115	.141	.170	.363	.636

Figure 4.B Variance of $\hat{\mu}(\gamma_o), \hat{\sigma}(\gamma_o)$ as Function of γ



4) For $\gamma = .333$ the variance is very large when γ_o is misspecified. When $\gamma = .05$, the variance when γ_o is misspecified is not significantly different from when γ_o is correctly specified.

Remarks on Tables 4.3 and 4.4 and Figures 4.C and 4.D

Tables 4.3 and Figure 4.C present the results of a study of $MSE(\hat{\mu}(\gamma_o))$ and $MSE(\hat{\sigma}(\gamma_o))$ for the sample size n = 20. Table 4.4 and Figure 4.D present analogous results for the sample size n = 100. The following remarks can be made:

1) For small sizes the variance term will dominate the bias term when computing MSE. As sample size increases, the effect decreases.

2) The curves for MSE look surprisingly like the curves for the variance of the estimators. Examination of Table 4.1 reveals that $[Bias(\hat{\mu}(\gamma_o))]^2$ is approximately equal to $nVar(\hat{\mu}(\gamma_o))$ and $[Bias(\hat{\sigma}(\gamma_o))]^2$ is approximately equal to $nVar(\hat{\sigma}(\gamma_o))$.

3) Since we are comparing biased and unbiased estimators of μ and σ, it is reasonable to compare the MSE of the estimators.

Table 4.3 MSE of $\hat{\mu}(\gamma_o)$, $\hat{\sigma}(\gamma_o)$ n = 20

	γ_o									
γ	.333	.250	.200	.167	.143	.125	.111	.100	.067	.050
.333	.004	.063	.279	.678	1.269	2.058	3.046	4.236	13.222	27.293
	.004	.077	.322	.758	1.390	2.222	3.256	4.492	13.717	28.035
.250	.031	.005	.048	.188	.433	.787	1.253	1.831	6.430	13.894
	.038	.005	.053	.204	.463	.833	1.317	1.913	6.610	14.176
.200	.082	.029	.005	.037	.134	.300	.537	.846	3.486	7.969
	.098	.032	.005	.039	.141	.314	.558	.875	3.563	8.097
.167	.139	.079	.025	.006	.030	.101	.221	.391	2.003	4.903
	.162	.086	.026	.005	.031	.104	.228	.402	2.040	4.969
.143	.193	.133	.066	.021	.006	.025	.080	.171	1.190	3.161
	.220	.143	.069	.021	.006	.025	.081	.175	1.208	3.197
.125	.244	.187	.114	.055	.018	.006	.021	.064	.708	2.083
	.274	.200	.120	.057	.018	.006	.021	.065	.718	2.104
.111	.290	.236	.162	.093	.046	.016	.006	.018	.420	1.400
	.321	.251	.169	.099	.047	.016	.006	.018	.424	1.412
.100	.332	.282	.208	.139	.082	.039	.014	.006	.242	.948
	.364	.298	.217	.143	.083	.040	.014	.006	.244	.955
.067	.488	.453	.390	.323	.259	.201	.150	.106	.006	.116
	.519	.470	.401	.330	.264	.204	.152	.107	.006	.117
.050	.587	.561	.509	.451	.393	.338	.286	.237	.063	.007
	.614	.577	.519	.458	.398	.341	.288	.239	.064	.006

Table 4.4 MSE of $\hat{\mu}(\gamma_o)$, $\hat{\sigma}(\gamma_o)$ n = 100

	γ_o									
γ	.333	.250	.200	.167	.143	.125	.111	.100	.067	.050
.333	.018	.086	.318	.738	1.356	2.179	3.206	4.441	13.743	28.281
	.021	.107	.372	.833	1.497	2.366	3.444	4.729	14.293	29.099
.250	.047	.024	.076	.227	.487	.859	1.346	1.949	6.715	14.422
	.052	.024	.083	.247	.522	.912	1.419	2.040	6.912	14.729
.200	.099	.045	.027	.066	.173	.351	.601	.926	3.672	8.308
	.112	.047	.026	.068	.181	.366	.625	.958	3.755	8.446
.167	.155	.092	.042	.028	.059	.139	.268	.449	2.136	5.143
	.175	.098	.042	.027	.060	.142	.276	.461	2.175	5.213
.143	.209	.145	.080	.039	.029	.055	.117	.216	1.291	3.342
	.233	.153	.082	.038	.028	.054	.118	.219	1.310	3.380
.125	.259	.197	.126	.070	.037	.030	.051	.100	.788	2.225
	.286	.208	.130	.071	.036	.029	.050	.110	.797	2.245
.111	.305	.245	.173	.109	.062	.036	.031	.048	.484	1.514
	.333	.259	.178	.111	.062	.034	.029	.047	.488	1.525
.100	.354	.290	.217	.150	.095	.056	.035	.031	.296	1.042
	.374	.304	.224	.153	.096	.056	.033	.030	.297	1.048
.067	.497	.458	.395	.329	.266	.210	.160	.118	.032	.161
	.526	.474	.405	.335	.270	.212	.161	.118	.031	.160
.050	.593	.564	.512	.454	.397	.343	.292	.245	.079	.033
	.619	.580	.521	.461	.402	.346	.294	.246	.078	.032

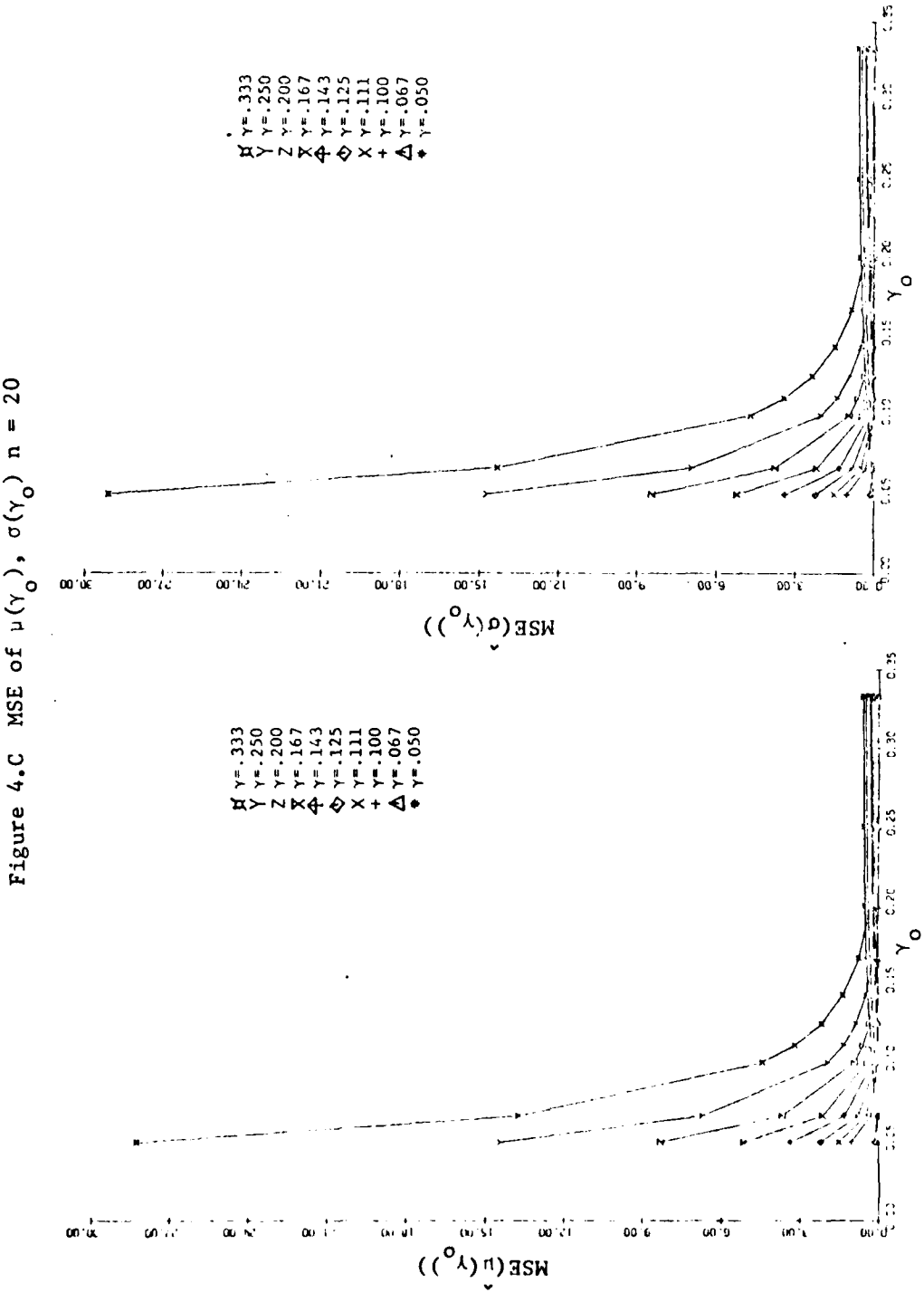


Figure 4.C MSE of $\hat{\mu}(\gamma_o)$, $\hat{\sigma}(\gamma_o)$ n = 20

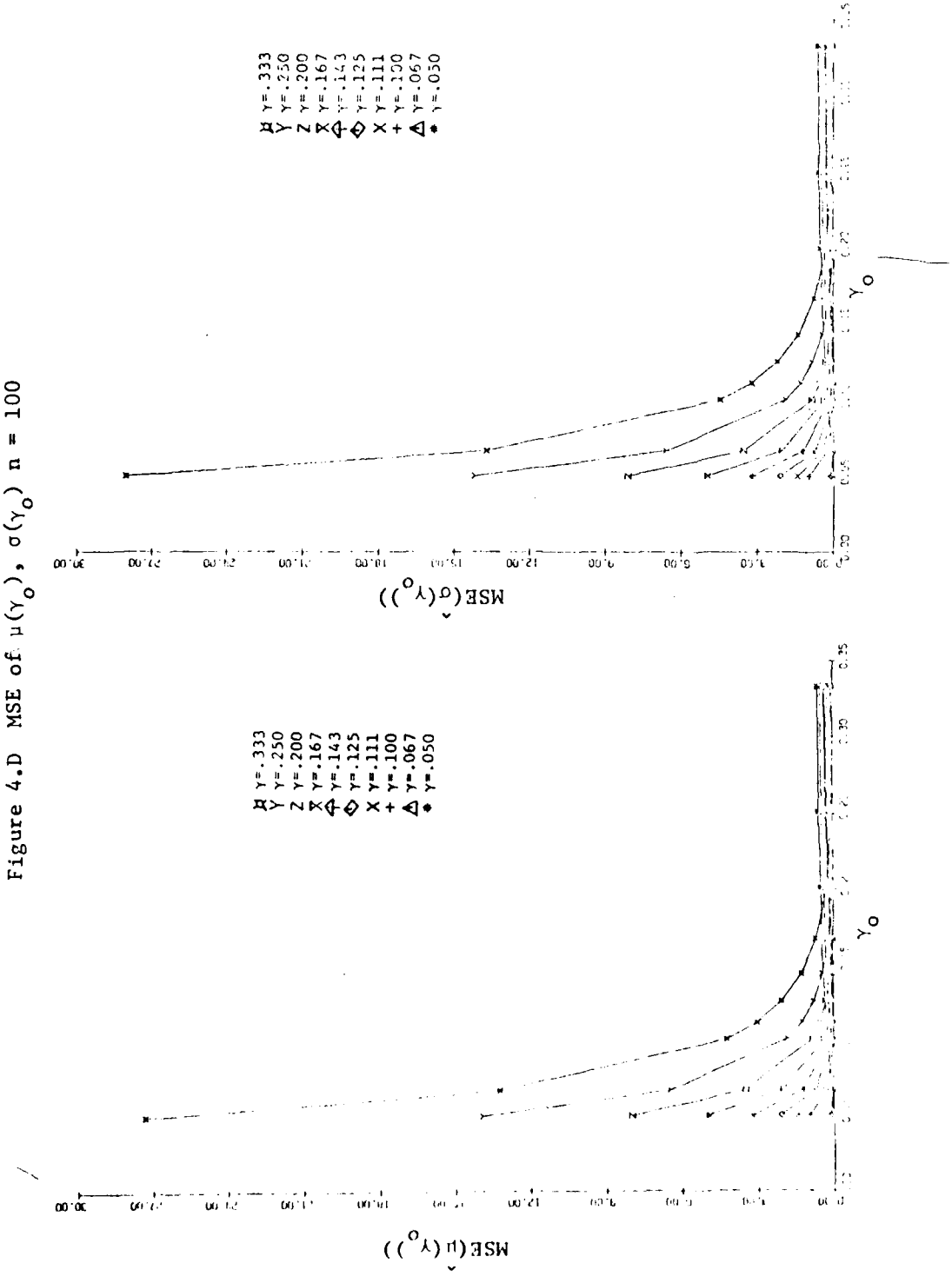


Figure 4.D MSE of $\hat{\mu}(Y_o)$, $\hat{\sigma}(Y_o)$ n = 100

4) When $\gamma = .333$, $MSE(\hat{\mu}(\gamma_o))$ and $MSE(\hat{\sigma}(\gamma_o))$ is very large when γ_o is misspecified. However when $\gamma = .05$, $MSE(\hat{\mu}(\gamma_o))$ and $MSE(\hat{\sigma}(\gamma_o))$ do not change significantly when γ_o is considerably misspecified.

Based on consideration of the bias, variance, and MSE of each estimator, the worst situation is to have data that is almost symmetric and to misspecify it as very skewed left. The consequences of misspecification are not severe when the data is skewed left. One will not do too badly if he uses estimators of μ and σ based on specifying $\gamma_o = .333$, .25, or .20 regardless of whether the distribution is skewed left or symmetric.

The techniques of determining γ investigated in Sections 3.2 and 3.3 seem to lead to a specification of γ that is in the range of the true value of γ. Thus estimators of μ and σ based on such a specification should yield reliable estimates of μ and σ.

5. QUANTILE REGRESSION AND COMPARISON FOR K SAMPLES

In the previous sections we described techniques to

1) identify Q_o or estimate γ (Section 3),

2) estimate μ and σ (Section 4)

in the model

$$Q(u) = \mu + \sigma\, Q_o(u) \quad .$$

The results of Section 3 have been generalized to the k sample problem. The results of Section 4 generalize directly to the estimation of location and scale parameters using independent samples from k populations.

Let us restate the model for k sample quantile regression. We assume

$$Q_i(u) = \mu_i + \sigma_i\, Q_o(u) \ , \ i = 1, \ldots k \ ,$$

where

$$\mu_i = \alpha_\mu + \beta_\mu X_i \ ,$$

$$\sigma_i = \alpha_\sigma + \beta_\sigma X_i \ ,$$

$Q_i(u)$ is the quantile function of the ith population, $Q_o(u)$ is an unknown quantile function or $Q_o(u) = (Q_o^*(u))^\gamma$ where $Q_o^*(u)$ is completely specified and γ is an unknown shape parameter common to the k populations, and X_i is a numerical characteristic of the ith

population. We further assume $X_1 \leq \ldots \leq X_k$ for convenience.

A commonly assumed quantile regression model is

$$Q_i(u) = A(u) + B(u)\, X_i$$

where A(u) and B(u) are unknown constants which depend on u. We desire estimators of A(u) and B(u) for a specified value of u.

Section 5.1 discusses the contributions of Brown and Mood (1950) and Hogg (1975) in the area of nonparametric quantile regression and the work of Griffiths and Willcox (1978) in parametric quantile regression. We also show the equivalence of the model of (5.1.1) and (5.1.2) to the model of (5.1.3) citing the work of Griffiths and Willcox (1978).

In Section 5.2 the generalized least squares technique is used to estimate α_μ, β_μ, α_σ, and β_σ. We state pertinent hypotheses about the regression parameters and provide test statistics for the hypotheses based on the asymptotic distribution of the estimated parameters.

Section 5.3 discusses how the regression technique of Section 5.2 can be applied to the k-sample comparison problem under certain restrictions. Test statistics for pertinent hypotheses about the location and/or scale parameters of the k populations are provided.

5.1 K Sample Quantile Regression

In this section we discuss the nonparametric technique of Hogg (1975) and the parametric approach of Griffiths and Willcox (1978) to estimate k sample quantile (percentile) relationships. The data for

the regression problem may consist of k random samples $\{Y_{i1}, \ldots, Y_{in_i}, i = 1, \ldots, k\}$ of the dependent variable together with the corresponding values $\{X_i, i = 1, \ldots, k\}$ of the independent variable or it may consist of a bivariate sample $\{(X_i, Y_i), i = 1, \ldots, n\}$.

Brown and Mood (1950) propose a nonparametric technique to estimate the median regression line for the model median $(Y|X) = \alpha + \beta X$ based on examining the residuals of the regression. They assume that for the bivariate sample $\{(X_i, Y_i), i = 1, \ldots, n\}$ the errors $Y_i - \alpha - \beta X_i$ have the same distribution for all X. They then estimate the median of the distribution of Y given X by

$$\tilde{\alpha} + \tilde{\beta} X \quad ,$$

where

$$\text{median } (Y_i - \tilde{\alpha} - \tilde{\beta}X_i) = 0 \quad X_i \leq \text{median}(X) \quad ,$$

$$\text{median } (Y_i - \tilde{\alpha} - \tilde{\beta}X_i) = 0 \quad X_i > \text{median}(X) \quad .$$

In words they split the sample into two subsamples of size n_1 and n_2, $n_i = n/2$, and then graphically find estimates $\tilde{\alpha}$, $\tilde{\beta}$ so that the median of the residuals $Y_i - \tilde{\alpha} - \tilde{\beta}X_i$ is zero for each batch.

Hogg (1975) modifies the technique of Brown and Mood in a natural way to estimate the pth percentile of the Y's. Hogg's model is

$$Y_p \doteq \Lambda(p) + B(p)X \quad ,$$

where Y_p denotes the pth percentile of the Y observations for a particular value of X. In terms of the quantile function we can write

$$Q_{Y|X}(p) = A(p) + B(p)X$$

where $Q_{Y|X}(\cdot)$ is the quantile function of Y given X and A(p) and B(p) are unknown constants which may vary with p. By examining the signs of the residuals, Hogg estimates the regression line of the pth percentile so that a fraction p of the data points are below the regression line. Hogg proposes statistics for testing

$$H_o: \quad Q_{Y|X}(u) = A_o(u) + B_o(u)X$$

where $A_o(u)$ and $B_o(u)$ are specified based on the binomial distribution of the number of observations below the hypothesized regression line. Several alternative procedures for splitting the data into more than two groups are also proposed .

Griffiths and Willcox (1978) assume the model

$$\frac{Q_{Y|X}(u)-\mu_X}{\sigma_X} = Q_o(u) \quad , \text{ for all } X \quad ,$$

where

$$\mu_X = \alpha_\mu + \beta_\mu X \quad ,$$

$$\sigma_X = \alpha_\sigma + \beta_\sigma X \quad .$$

This is equivalent to assuming

$$Q_{Y|X}(u) = A(u) + B(u)X \quad ,$$

i.e. a linear regression model, where

$$A(u) = \alpha_\mu + \alpha_\sigma Q_o(u) \quad ,$$

$$B(u) = \beta_\mu + \beta_\sigma Q_o(u) \quad .$$

They then estimate α_μ, β_μ, α_σ, β_σ using iterative maximum likelihood procedures on the weighted residuals

$$[Q_{Y|X}(u) - (\alpha_\mu + \beta_\mu X)]/(\alpha_\sigma + \beta_\sigma X) \quad .$$

By weighting the estimates $\hat{\alpha}_\mu$, $\hat{\beta}_\mu$, $\hat{\alpha}_\sigma$, $\hat{\beta}_\sigma$ using their estimated variance matrix, point estimates or interval estimates of A(u) and B(u) can be computed. The authors use $Q_o(u) = \Phi^{-1}(u)$. The likelihood equations do not have a closed-form solution, and Griffiths and Willcox use linear programming methods to determine optimal values for the parameters. They state that an advantage in using a parametric model for the data is a gain in precision and efficiency.

In the next section we describe a parametric approach to solving the k sample quantile regression problem based on a quantile function approach. The procedure is more general than that of Griffiths and Willcox yet incorporates parametric assumptions which give it certain advantages over the nonparametric technique of Hogg.

5.2 A Quantile Function Procedure for K Sample Quantile Regression

In this section we generalize the model of Griffiths and Willcox (1978) to allow any $Q_o(u)$. We assume

$$Q_{Y|X}(u) = \mu_X + \sigma_X Q_o(u)$$

or more specifically

$$Q_i(u) = \mu_i + \sigma_i Q_o(u) \ , \ i=1, \ldots, k \ , \qquad (5.2.1)$$

where

$$\begin{aligned} \mu_i &= \alpha_\mu + \beta_\mu X_i \\ \sigma_i &= \alpha_\sigma + \beta_\sigma X_i \ , \ i = 1, \ldots, k \ . \end{aligned} \qquad (5.2.2)$$

Using the estimates $\hat{\mu}_i$, $\hat{\sigma}_i$ based on LCOS (Section 4), we obtain estimates of α_μ, β_μ, α_σ, β_σ using generalized least squares.

The first step is to identify the Q_o function common to all k populations using the techniques of Section 3.1 and 3.3 or, if appropriate, to estimate the shape parameter γ of the specified Q_o^* using the techniques of Section 3.2 and 3.3.

From each of the random samples $\{Y_{i1}, \ldots, Y_{in_i}, i=1, \ldots,k\}$ one forms estimates $\hat{\mu}_i$ and $\hat{\sigma}_i$ of μ_i and σ_i using optimal or asymptotically optimal LCOS.

Theorem 5.2.1: Assuming that the standard conditions for the validity of the Cramer-Rao bounds are satisfied and that the spacings $\{ 0, u_1, \ldots, u_r, 1 \}$ satisfy

$$\max_j (u_j - u_{j-1}) \to 0 \text{ as } r \to \infty \quad ,$$

then

$$\sqrt{n}\left(\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix} - \begin{pmatrix}\mu\\ \sigma\end{pmatrix}\right) \xrightarrow{d} N_2 \left(\underset{\sim}{0}_2, \sigma^2 I^{-1}(\underset{\sim}{\theta})\right) \text{ as } r \to \infty$$

where $\hat{\mu}$, $\hat{\sigma}$ are computed using the optimal or asymptotically optimal LCOS based on r order statistics and $I(\underset{\sim}{\theta})$ is the Fisher information matrix of (μ, σ) defined by (4.1.6).

Proof: (nonrigorous)

The asymptotic zero mean normality of $\sqrt{n}\left(\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix} - \begin{pmatrix}\mu\\ \sigma\end{pmatrix}\right)$ follows from Theorem 2.1.4 and the fact that $\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix}$ are ABLUE's for $\begin{pmatrix}\mu\\ \sigma\end{pmatrix}$.

The variance of $\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix}$ is given by

$$V\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix} = \sigma^2 \begin{pmatrix}K_{11} & K_{12}\\ K_{12} & K_{22}\end{pmatrix}^{-1}$$

(see Ogawa 1951 or Eubank 1979) where K_{11}, K_{12}, and K_{22} are defined by (4.1.1), (4.1.2), and (4.1.3). It is sufficient to show that $\begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix}$ converges componentwise to $I(\underset{\sim}{\theta})$ as $r \to \infty$.

Consider K_{11} defined by

$$K_{11} = \sum_{j=1}^{r+1} \frac{(f_o Q_o(u_j) - f_o Q_o(u_{j-1}))^2}{u_j - u_{j-1}} .$$

Using the Mean Value Theorem,

$$f_o Q_o(u_j) = f_o Q_o(u_{j-1}) + (u_j - u_{j-1}) f_o Q_o'(u_j^*)$$

so that

$$\frac{f_o Q_o(u_j) - f_o Q_o(u_{j-1})}{u_j - u_{j-1}} = f_o Q_o'(u_j^*) ,$$

where $u_{j-1} \le u_j^* \le u_j$, $u_0 = 0$, $u_{r+1} = 1$. Then we can write the Riemann integral

$$I_{11}(\underset{\sim}{\theta}) = \int_0^1 (f_o Q_o'(u))^2 \, du$$

$$= \lim_{r \to \infty} \sum_{j=1}^{r+1} \left(f_o Q_o'(u_j^*) \right)^2 (u_j - u_{j-1})$$

as

$$I_{11}(\underset{\sim}{\theta}) = \lim_{r \to \infty} \sum_{j=1}^{r+1} \frac{(f_o Q_o(u_j) - f_o Q_o(u_{j-1}))^2}{u_j - u_{j-1}} .$$

The convergence of K_{22} to $I_{22}(\underset{\sim}{\theta})$ and K_{12} to $I_{12}(\underset{\sim}{\theta})$ follow analogously.

Remarks on Theorem 5.2.1:

1) A rigorous proof would entail examining the limits $n \to \infty$ and $r \to \infty$ more closely as well as the asymptotic normality of $(\hat{\mu}, \hat{\sigma})$ (see Chernoff, Gastwirth, and Johns 1967, and Stigler 1974). Theorem 2.1.4 holds for fixed $u_1, \ldots, u_r$ but here we let $r \to \infty$.
2) It seems clear that the asymptotically optimal spacings generated by Eubank (1979) satisfy the conditions of Theorem 5.2.1 since $H(u)$ and $H^{-1}(u)$ are both defined on $[0,1]$. One should substitute $H^{-1}(\frac{j}{r+1})$ for u_j in the expressions for K_{11}, K_{12}, and K_{22} to get the variance of Eubank's estimators.
3) It is not clear that the optimal spacings of Ogawa (1950) satisfy the conditions of Theorem 5.2.1. Examination of the ARE of the estimators of μ and σ using Ogawa's formulation suggest that $ARE(\hat{\mu}, \hat{\sigma}) \to 1$ as $r \to \infty$. This seems to indicate that the conditions on the spacings are satisfied.

Corollary 5.2.1: When there are k independent samples, the estimators $\hat{\mu}_i$ and $\hat{\sigma}_i$ based on LCOS from a random sample of size n_i from population i satisfy

$$\sqrt{n_i}\left(\begin{pmatrix}\hat{\mu}_i\\ \hat{\sigma}_i\end{pmatrix} - \begin{pmatrix}\mu_i\\ \sigma_i\end{pmatrix}\right) \xrightarrow{d} N_2(\underset{\sim}{0}_2,\ \sigma_i^2\, I^{-1}(\underset{\sim}{\theta})) \text{ as } r \to \infty$$

and the k distributions are independent.

Theorem 5.2.2: a) When there are k populations satisfying the model of (5.2.1) and (5.2.2), then ABLUE's $(\hat{\alpha}_\mu,\ \hat{\beta}_\mu,\ \hat{\alpha}_\sigma,\ \hat{\beta}_\sigma)$ of $(\alpha_\mu,\ \beta_\mu,\ \alpha_\sigma,\ \beta_\sigma)$ are given by

$$\begin{pmatrix}\hat{\alpha}_\mu\\ \hat{\beta}_\mu\\ \hat{\alpha}_\sigma\\ \hat{\beta}_\sigma\end{pmatrix} = I^{-1}(\underset{\sim}{\theta}) \otimes \left[\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} W_i\right]^{-1} \left\{\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2}\left[I_2 \otimes \begin{pmatrix}1\\ X_i\end{pmatrix}\right] I(\underset{\sim}{\theta}) \begin{pmatrix}\hat{\mu}_i\\ \hat{\sigma}_i\end{pmatrix}\right\}, \tag{5.2.3}$$

where $I(\underset{\sim}{\theta})$ is the Fisher information matrix of (μ,σ) defined by (4.1.6), I_n is the n dimensional identity matrix, and the Kronecker product, $A \otimes B$ (Rao 1973, p. 29), of an n x m matrix $A = (A_{ij})$ and an r x s matrix $B = (B_{ij})$ is the nr x ms matrix

$$A \otimes B = \begin{bmatrix} a_{11} B & a_{12} B & \cdots & a_{1m} B \\ \vdots & & & \\ a_{n1} B & \cdots & & a_{nm} B \end{bmatrix},$$

and

$$W_i = \begin{bmatrix} 1 & X_i \\ X_i & X_i^2 \end{bmatrix}, \quad i = 1, \ldots, k .$$

b)

$$V^{-\frac{1}{2}} \left(\begin{pmatrix} \hat{\alpha}_\mu \\ \hat{\beta}_\mu \\ \hat{\alpha}_\sigma \\ \hat{\beta}_\sigma \end{pmatrix} - \begin{pmatrix} \alpha_\mu \\ \beta_\mu \\ \alpha_\sigma \\ \beta_\sigma \end{pmatrix} \right) \xrightarrow{d} N_4 \left(\underset{\sim}{0}_4 , I_4 \right) ,$$

where

$$V = I^{-1}(\underset{\sim}{\theta}) \otimes \begin{bmatrix} \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} W_i \end{bmatrix}^{-1} \qquad (5.2.4)$$

and $V^{-\frac{1}{2}} V^{-\frac{1}{2}} = V^{-1}$.

Proof:

Using model (5.2.2) we can write

$$\begin{pmatrix} \hat{\mu}_i \\ \hat{\sigma}_i \end{pmatrix} = \begin{bmatrix} 1 & X_i & 0 & 0 \\ 0 & 0 & 1 & X_i \end{bmatrix} \begin{pmatrix} \alpha_\mu \\ \beta_\mu \\ \alpha_\sigma \\ \beta_\sigma \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} , \quad i = 1, \ldots, k ,$$

or

$$\begin{pmatrix} \hat{\mu}_i \\ \hat{\sigma}_i \end{pmatrix} = I_2 \otimes (1 \; X_i) \begin{pmatrix} \alpha_\mu \\ \beta_\mu \\ \alpha_\sigma \\ \beta_\sigma \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} , \quad i = 1, \ldots, k ,$$

where

$$\mathrm{Var} \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} = \frac{\sigma_i^2}{n_i} \; I^{-1}(\underset{\sim}{\theta}) .$$

Using the asymptotic normality of $\begin{pmatrix} \hat{\mu}_i \\ \hat{\sigma}_i \end{pmatrix}$ (Theorem 5.2.1), the observations $\begin{pmatrix} \hat{\mu}_i \\ \hat{\sigma}_i \end{pmatrix}$ approximately fit the framework of the Gauss-Markov Theorem (see Rao 1973,pp.544-546) and we can form the ABLUE's $(\hat{\alpha}_\mu, \hat{\beta}_\mu, \hat{\alpha}_\sigma, \hat{\beta}_\sigma)$ of $(\alpha_\mu, \beta_\mu, \alpha_\sigma, \beta_\sigma)$ given by (5.2.3) which have asymptotic variance given by (5.2.4).

Remarks on Theorem 5.2.2:

1) Notice that σ_i^2 is an unknown parameter so that usually an iterative estimation scheme is in order. However, analogous to the treatment of σ^2 in the continuous parameter time series regression model of (4.2.2), we can treat σ_i^2 as the scale parameter of a Brownian bridge process (see Parzen 1979a). Hence under the assumption of (5.2.1), i.e. that $Q_i(u)$ is a location-scale shift of some $Q_o(u)$, an "independent" estimator of σ_i is provided by $\tilde{\sigma}_{o_i}$ where

$$\tilde{\sigma}_{o_i} = \int_0^1 f_o Q_o(u) \tilde{q}_i(u) du \quad .$$

This is the k-sample analog of $\tilde{\sigma}_o$ defined in Section 3.1. The estimator is consistent for σ_i when (5.2.1) is true. Consequently we compute the estimators

$$\begin{pmatrix} \hat{\alpha}_\mu \\ \hat{\beta}_\mu \\ \hat{\alpha}_\sigma \\ \hat{\beta}_\sigma \end{pmatrix} = I^{-1}(\underset{\sim}{\theta}) \otimes \left[\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} W_i\right]^{-1} \left\{\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} \left[I_2 \otimes \begin{pmatrix} 1 \\ X_i \end{pmatrix}\right] I(\underset{\sim}{\theta}) \begin{pmatrix} \hat{\mu}_i \\ \hat{\sigma}_i \end{pmatrix}\right\} \tag{5.2.5}$$

with estimated variance

$$\hat{V} = I^{-1}(\underset{\sim}{\theta}) \otimes \left(\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} W_i\right)^{-1} \tag{5.2.6}$$

3) Programs to implement the estimation techniques are available. Subroutine KSAM forms quantile-box plots and uses the goodness-of-fit techniques of Section 3.3 to determine the distributional shape of the k samples. Subroutines QTOLS, QTOLSC, and QTOLSW compute estimates $\hat{\mu}_i$ and $\hat{\sigma}_i$ of μ_i and σ_i for a specified Q_o function using LCOS. Subroutine LSTOAB estimates the coefficients $(\alpha_\mu, \beta_\mu, \alpha_\sigma, \beta_\sigma)$ and their variance using (5.2.5) and (5.2.6) based on the k pairs of observations $(\hat{\mu}_i, \hat{\sigma}_i)$, i = 1, ..., k. Listings of the subroutines are on file at the Institute of Statistics, Texas A&M University.

4) Model (5.2.2) has been used for simplicity. A general parametric model relating μ_i and σ_i to X_i is

$$\mu_i = f_\mu(X_i, \underset{\sim}{\theta}_\mu)$$

$$\sigma_i = f_\sigma(X_i, \underset{\sim}{\theta}_\sigma), \quad i = 1, \ldots, k \ .$$

Scatter diagrams of μ_i and σ_i vs X_i will help determine appropriate $f_\mu(\cdot\ ,\ \cdot)$ and $f_\sigma(\cdot\ ,\ \cdot)$ functions.

The final step in the k-sample quantile regression problem is to estimate the parameters A(u) and B(u) in

$$Q_i(u) = A(u) + B(u)\ X_i\ ,\ i = 1,\ \ldots,\ k\ .$$

By making the substitution

$$A(u) = \alpha_\mu + \alpha_\sigma\ Q_o(u)\ ,$$

$$B(u) = \beta_\mu + \beta_\sigma\ Q_o(u)\ ,$$

we obtain the estimators

$$\begin{aligned} \hat{A}(u) &= \hat{\alpha}_\mu + \hat{\alpha}_\sigma\ Q_o(u)\ , \\ \hat{B}(u) &= \hat{\beta}_\mu + \hat{\beta}_\sigma\ Q_o(u)\ . \end{aligned} \qquad (5.2.7)$$

A significant advantage in this estimation scheme over other methods is that one need not use sophisticated methods to estimate A(u) and B(u) for each value of u for which a regression line is desired. One can simply substitute the appropriate value of $Q_o(u)$ in (5.2.7).

Hypothesis testing procedures:

The first hypothesis of interest is

$$H_o:\ Q_i(u) = \mu_i + \sigma_i\ Q_o(u)\ ,\ i = 1,\ \ldots,\ k\ .$$

This hypothesis examines the adequacy of the model $Q_i(u) = \mu_i + \sigma_i Q_o(u)$.

To test this hypothesis we propose the GOF procedures outlined in Section 3.3.

One might also wish to examine the adequacy of the linear model (5.2.2) for μ_i and σ_i. Scatter diagrams of $\hat{\mu}_i$ vs. X_i and $\hat{\sigma}_i$ vs. X_i provide a quick graphic technique to check the linear relationship between the estimated parameters and the X values.

A hypothesis which states that there is no linear relationship between the quantiles of Y and X is

$$H_o: \beta_\mu = \beta_\sigma = 0 \quad .$$

To test this hypothesis one could use the joint asymptotic distribution of $\hat{\beta}_\mu$ and $\hat{\beta}_\sigma$ given by Theorem 5.2.2 and form the test statistic

$$X^2 = (\hat{\beta}_\mu, \hat{\beta}_\sigma)[\widehat{\mathrm{Var}}(\hat{\beta}_\mu, \hat{\beta}_\sigma)]^{-1} \begin{pmatrix} \hat{\beta}_\mu \\ \hat{\beta}_\sigma \end{pmatrix}$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_\mu, \hat{\beta}_\sigma)$ consists of the appropriate elements of the estimated variance matrix given by (5.2.6). Under H_o, X^2 has an asymptotic χ^2 distribution with two degrees of freedom. Large values of X^2 indicate departure from H_o.

Other hypotheses involving $\alpha_\mu, \beta_\mu, \alpha_\sigma, \beta_\sigma$ can be tested using the asymptotic normality of $(\hat{\alpha}_\mu, \hat{\beta}_\mu, \hat{\alpha}_\sigma, \hat{\beta}_\sigma)$. Some of these hypotheses are discussed in the next section.

5.3 The K-Sample Comparison Problem

The k-sample comparison problem is defined to be the estimation and comparison of the location and/or scale parameters of k populations based on samples $\{Y_{i1}, \ldots, Y_{in_i}, i = 1, \ldots, k\}$. Usually one assumes

$$F_i(y) = F_o\left(\frac{y-\mu_i}{\sigma_i}\right)$$

or

$$Q_i(u) = \mu_i + \sigma_i Q_o(u) \ , \ i = 1, \ldots, k \ ,$$

where μ_i and σ_i are the location and scale parameters respectively of the ith population and F_o and Q_o are completely specified. There are a multitude of parametric and nonparametric procedures available to compare the μ_i's or the σ_i's.

If one records some numerical characteristic, X_i, of the ith population, e.g. treatment level, one can specify a relationship between (μ_i, σ_i) and X_i such as

$$\begin{aligned} \mu_i &= \alpha_\mu + \beta_\mu X_i \ , \\ \sigma_i &= \alpha_\sigma + \beta_\sigma X_i \ . \end{aligned} \qquad (5.3.1)$$

Thus the estimation procedures of Section 5.2 are also appropriate for a particular type of location and scale comparison problem.

A hypothesis which examines the equality of the k location parameters is

$$H_\mu: \mu_1 = \ldots = \mu_k \quad .$$

Under the model (5.3.1) this hypothesis is equivalent to

$$H_\mu: \beta_\mu = 0 \quad .$$

In Section 5.2 we state the asymptotic variance of $\hat{\beta}_\mu$ from which we can form the test statistic

$$z_\mu = \hat{\beta}_\mu / (\hat{\text{Var}}(\hat{\beta}_\mu))^{\frac{1}{2}}$$

where $\hat{\text{Var}}(\hat{\beta}_\mu)$ is the appropriate element of (5.2.6). Under H_μ, z_μ has an asymptotic N(0, 1) distribution. For the alternative $H_\sigma: \beta \neq 0$, one rejects H_μ at level α if $| z_\mu | > \Phi^{-1}(1 - \alpha/2)$. The test is not appropriate, however, for the general alternative

$$H_\alpha: \text{not all } \mu_i \text{ are equal.}$$

A hypothesis which states the equality of the k scale parameters is

$$H_\sigma: \sigma_1 = \ldots = \sigma_k$$

which under model (5.3.1) is equivalent to

$$H_\sigma: \beta_\sigma = 0 \quad .$$

Analogous to the test statistic for H_μ, we can form the test statistic, z_σ, for H_σ defined by

$$z_\sigma = \hat{\beta}_\sigma / (\widehat{\mathrm{Var}}(\hat{\beta}_\sigma))^{\frac{1}{2}}$$

which has an asymptotic N(0, 1) distribution under H_σ.

One might also wish to test simultaneously the equality of the μ_i's and the σ_i's, i.e. test

$$H_o: \mu_1 = \ldots = \mu_k \text{ and } \sigma_1 = \ldots = \sigma_k$$

or

$$H_o: \beta_\mu = \beta_\sigma = 0 .$$

To test this hypothesis use the test statistic X^2 of Section 5.2 defined by

$$X^2 = (\hat{\beta}_\mu, \hat{\beta}_\sigma)[\widehat{\mathrm{Var}}\ (\hat{\beta}_\mu, \hat{\beta}_\sigma)]^{-1} \begin{pmatrix} \hat{\beta}_\mu \\ \hat{\beta}_\sigma \end{pmatrix}$$

which under H_o has an asymptotic χ^2 distribution with two degrees of freedom.

The advantages of this comparison procedure are:

1) There are no restrictions on Q_o (e.g. $Q_o = \phi^{-1}$).
2) One can accommodate heterogeneity of the σ_i's (or μ_i's) when comparing the μ_i's (or σ_i's). The standard ANOVA procedure for the comparison of location parameters

assumes $\sigma_1 = \ldots = \sigma_k$ and $Q_o = \phi^{-1}$. The resulting F statistic performs fairly well when either of the assumptions are violated but its performance worsens when both assumptions are violated particularly if the n_i's vary considerably (Box 1954) .

3) In many situations (e.g. the Weibull distribution) the location parameter is a threshhold value and not a measure of central tendency. The mean and median of the distribution will depend on scale and shape parameters as well as the location parameter. Thus a procedure based on comparing sample means or medians seems inappropriate for comparing location parameters.

A substantial disadvantage of the procedure is that location and scale comparisons based on the model of (5.3.1) have very low power against general alternatives.

6. EXAMPLES

In this section two published data sets are analyzed using the techniques of Sections 3 through 5. The results of the analyses are compared to results of other investigations of the data. Computing programs to implement the techniques of data analysis described in this dissertation have been developed by the author. The programs make use of subroutines which implement the nonparametric data modeling techniques of Parzen (1979). All computing was performed on an AMDAHL 470V/6 computer at Texas A&M University.

6.1 Professors' Salary Example

Hogg (1975), Griffiths and Willcox (1978), and Angers (1979) investigate data consisting of the salaries of 96 professors at a major university as a function of their years in service. Each of the investigators estimate linear percentile lines for p = .25, .50, .75. The techniques of Hogg (1975) and of Griffiths and Willcox (1978) have been described in Section 5.1. The approach of Angers (1979) is to use grafted polynomials, a nonparametric technique, where the curves for the 75th and 25th percentiles are restricted to be symmetric about the curve for the 50th percentile. He uses linear percentile regression curves. Table 6.1 summarizes the estimated quantile regression coefficients obtained by each author.

Table 6.1 Estimated Parameters for Professors' Salary Data.

	$\hat{A}(.25)$	$\hat{B}(.25)$	$\hat{A}(.50)$	$\hat{B}(.50)$	$\hat{A}(.75)$	$\hat{B}(.75)$
Hogg(1975)	18.8	.300	20.0	.485	21.5	.625
Griffiths & Willcox(1979)	17.50	.40	19.15	.48	20.81	.56
Angers(1979)	18.173	.331	19.646	.478	21.119	.625

The data are presented in Figure 6.A.

Griffiths and Willcox state:

> "There is no clear evidence in the data of departures from normality by way of either highly skewed or heavily tailed residual distributions. There is, however, a trend to increasing spread ..."

However, generally with salary data one would expect the data to be skewed right particularly when there are few years in service. There seem to be several outliers in the data when there are few years in service. While one might expect increasing spread of the salary distribution as years in service increases, the increase in spread evidenced by this data does not seem substantial.

What one would like to detect is how the quantiles behave as a function of years in service. One would like to determine which of the potential curves in Figure 6.B represents the relationship between salary quantiles and years in service and to estimate the unknown parameters of the quantile regression function.

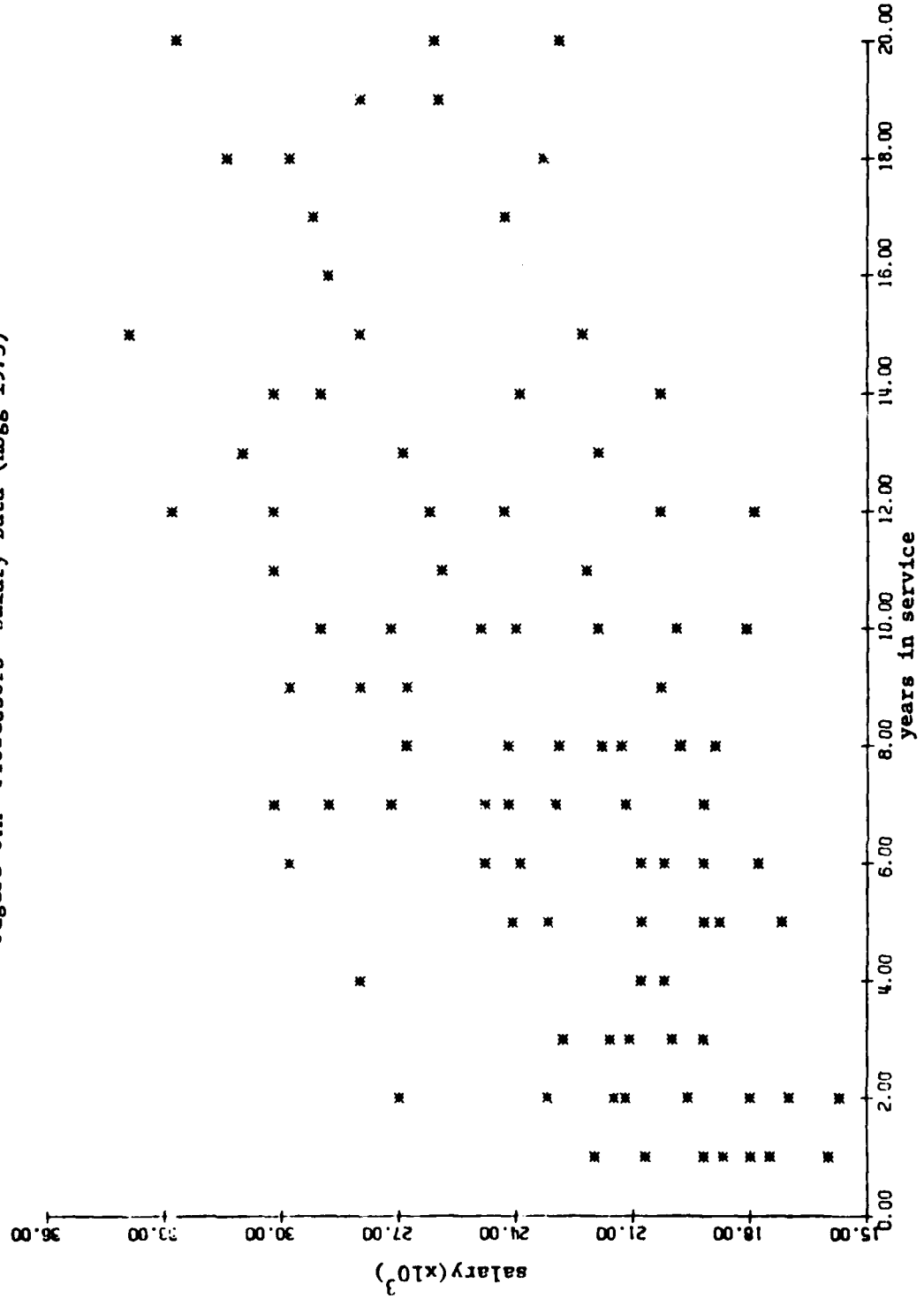
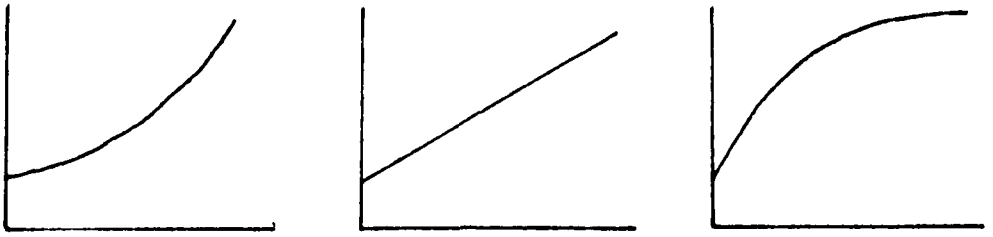


Figure 6.A Professors' Salary Data (Hogg 1975)

Figure 6.B Possible Salary Quantile Regression Curves



Since the sample sizes are quite small for each value of X, in order to use our quantile regression technique, it is necessary to repartition the data by pooling homogeneous samples. Tukey (1977) suggests that one way to partition Y observations when X is a random variable is to use selected quantiles of X. In this study three methods of partitioning the data were investigated:

1) pooling the data into four year intervals;
2) pool the data into five year intervals;
3) pool the data using similar midrange values which resulted in five samples representing 3, 3, 4, 4 and 6 years of service respectively.

It was found that pooling in four samples each representing five years of service, was most satisfactory for this study. It seems that there is a jump in salary after five years of service. Another method of partitioning the data is to pool the data into overlapping samples. However this technique violates the assumption of k independent samples.

Based on pooling the data into four samples of five years each, we shall describe each stage of the analysis. We use the mean value

of X within each sample as the value of X_i. Thus $X_1 = 3$, $X_2 = 8$, $X_3 = 13$, $X_4 = 18$.

Stage 1: The quantile-box plots of all four samples are given in Figure 6.C. The shift in location is very evident. The shapes of the distributions seem compatible but all of the plots show varying degrees of skewness. However all of the sample sizes are relatively mall and it is difficult to identify incompatible shapes from the quantile-box plots. The plots do suggest that one should test the goodness-of-fit of the data to symmetric and slightly skewed Q_o functions.

Using the technique of Section 3.3 we test the goodness-of-fit of the data to the normal, logistic, and Weibull ($\gamma = .333, .250, .20$) distributions.

By specifying $Q_o(u) = \Phi^{-1}(u)$ (normal distribution) we obtain the quantile-box plot of Figure 6.D for the pooled transformed data. The plot is not incompatible with a normal shape except in the tails. Figure 6.E is a plot of $\tilde{D}(u)$, the raw transformation distribution function. The line $D(u) = u$ has also been superimposed on the figure. Serious departures from the line $D(u) = u$ are not obvious. The value of $\tilde{\rho}(1)$ is .0206. Under H_o, $2n\tilde{\rho}(v)$, $v \neq 0$ has an asymptotic χ^2 distribution with two degrees of freedom. The .05 critical value of a χ^2_2 is 5.99. The value of $2n\tilde{\rho}(1)$ is 3.95 where $n = 96$ and $\tilde{\rho}(1) = .0206$. This is also evidence that the normal distribution is compatible with the data. Finally CAT selects an

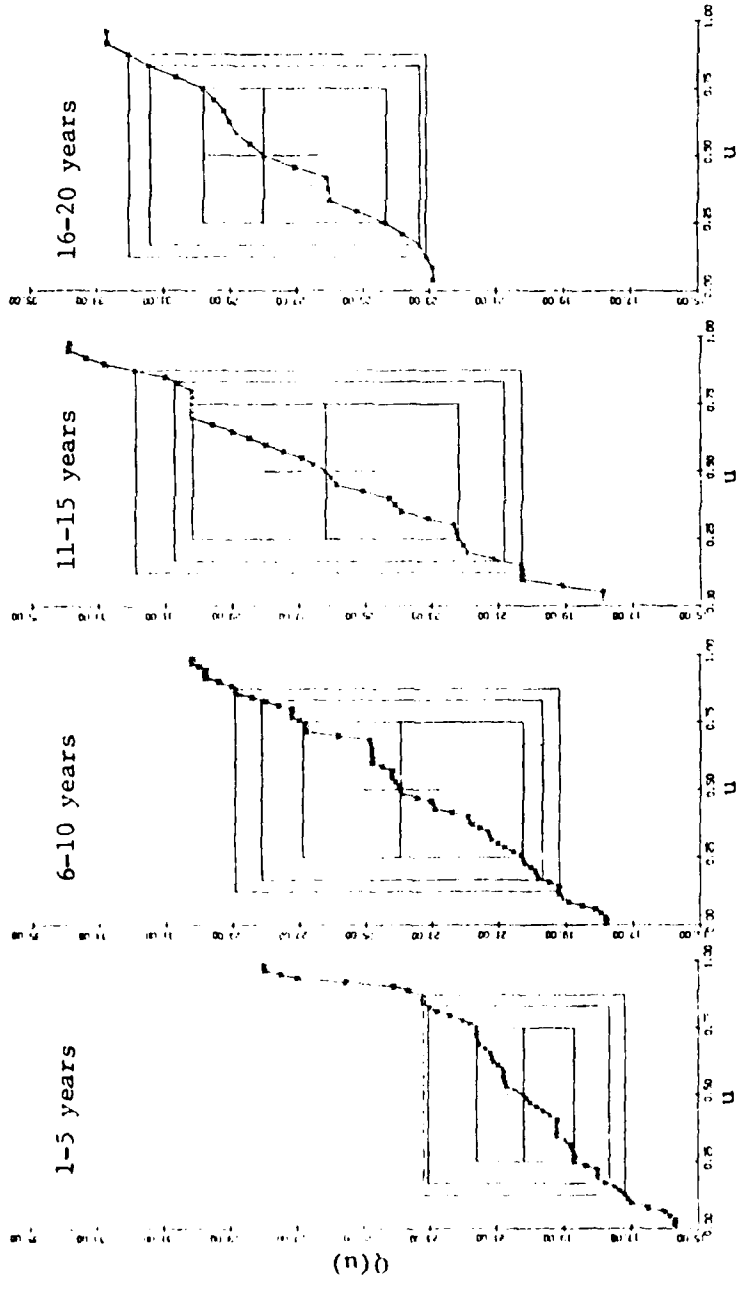


Figure 6.3 Professors' Salary Data;Quantile-Box Plots of Four Samples.

Figure 6.D Professors' Salary Data;Quantile-Box Plot of Pooled Transformed Data, Normal case

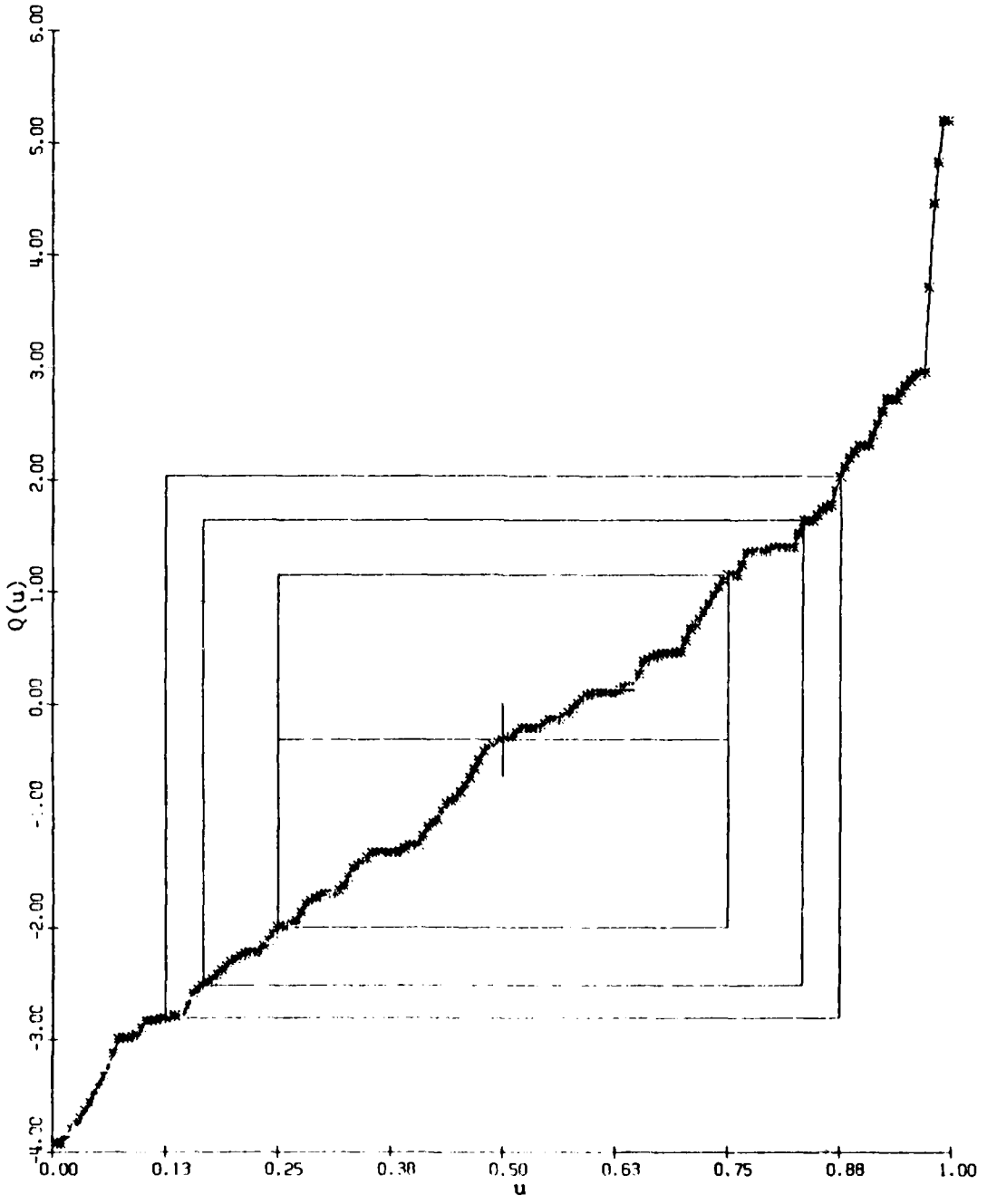
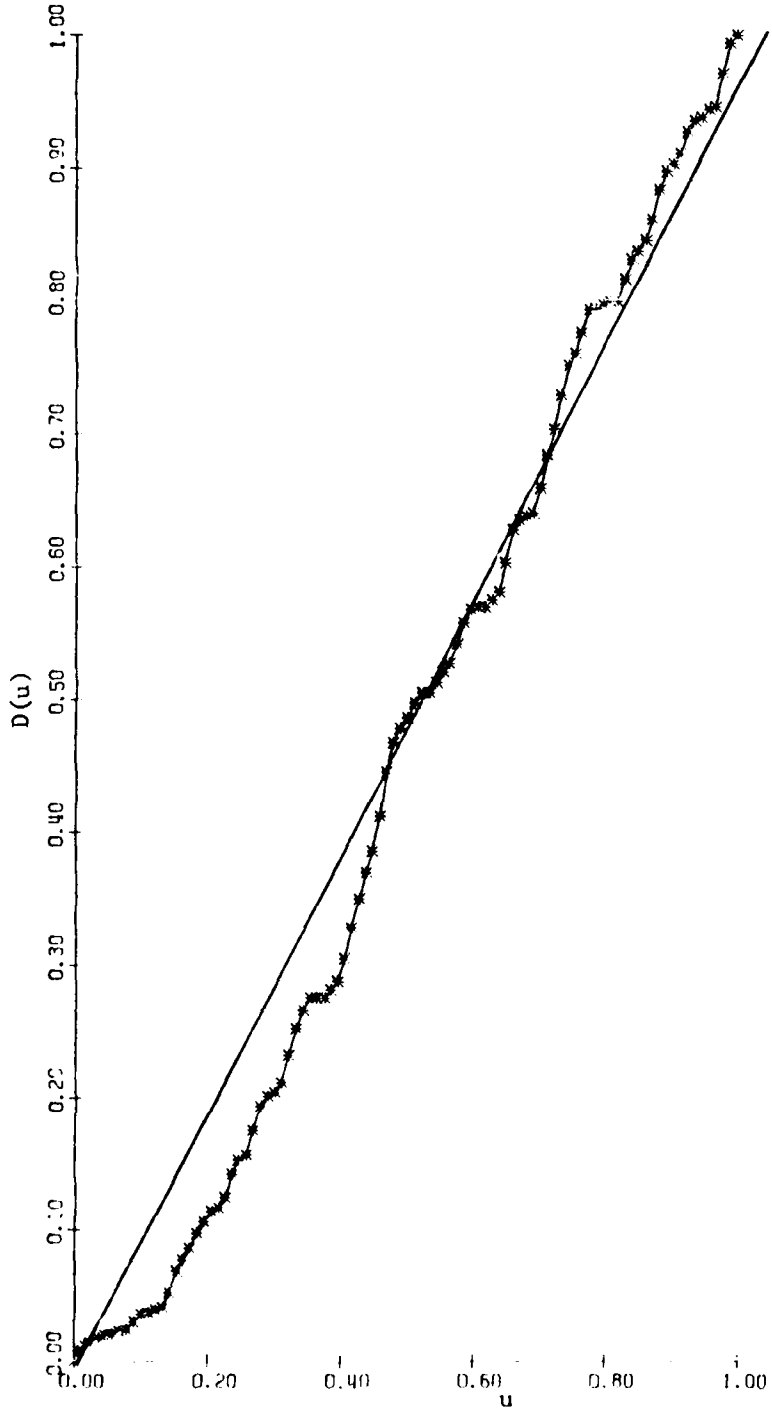
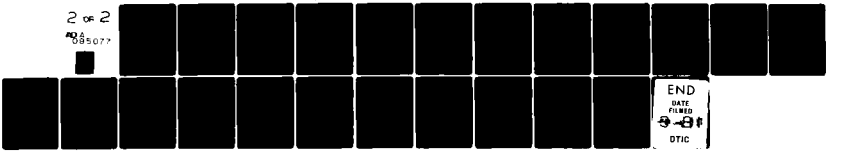


Figure 6.E Professors' Salary Data; The Function $\tilde{D}(u)$, Normal case



AD-A095 077 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
A QUANTILE FUNCTION APPROACH TO THE K-SAMPLE QUANTILE REGRESSIO--ETC(U)
NOV 80 J M WHITE DAAG29-80-C-0070
UNCLASSIFIED TR-B-4 ARO-16992.4-M NL

2 OF 2
AD A095077



END
DATE FILMED
3-81
DTIC

optimal order of zero which is consistent with the other diagnostics in failing to reject a normal distribution for the data. Based on this stage of analysis we conclude that $Q_i(u) = \mu_i + \sigma_i \Phi^{-1}(u)$, $i = 1, 2, 3, 4$. Using a consensus of the diagnostics, the other distributions, i.e. logistic and Weibull ($\gamma = .333, .250, .20$) , are not as compatible with the data as the normal distribution.

Stage 2: We compute estimates $\hat{\mu}_i$, $\hat{\sigma}_i$ of μ_i, σ_i using LCOS based on the normal distribution using the asymptotically optimal coefficients and spacings ($r = 7$) of Eubank (1979). Figure 6.F plots $\hat{\mu}_i$ and $\hat{\sigma}_i$ vs X_i. The figure suggests a linear model for the $\hat{\mu}_i$'s but it appears that there is no definite trend for the $\hat{\sigma}_i$'s. However we shall attempt to fit the linear model of (5.2.1) for both μ_i and σ_i in Stage 3.

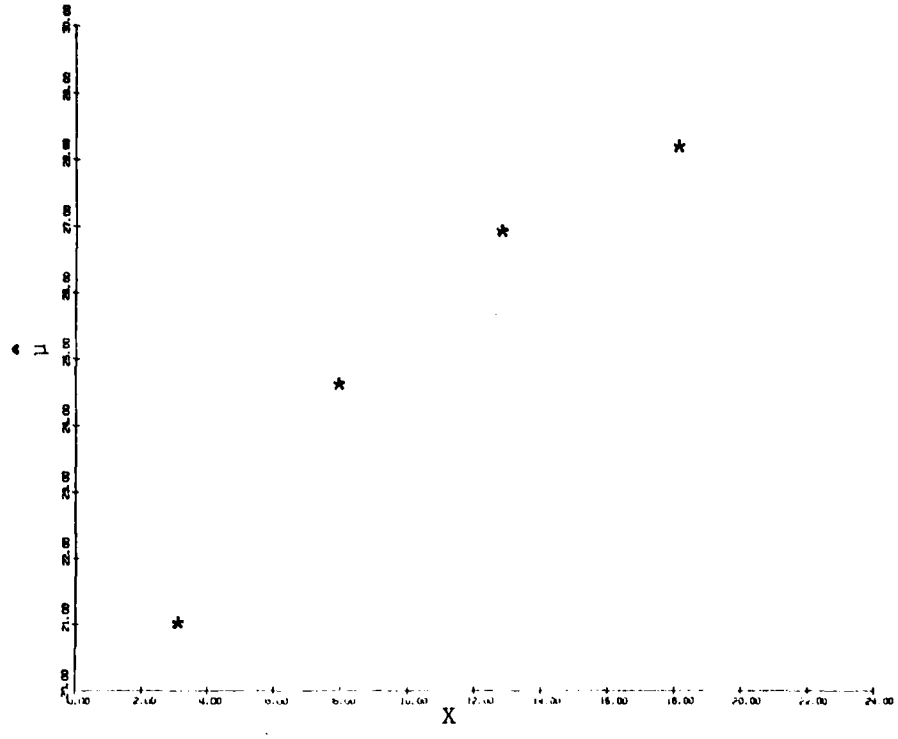
Stage 3: We use generalized least squares to obtain the following fitted regression lines for μ_i and σ_i:

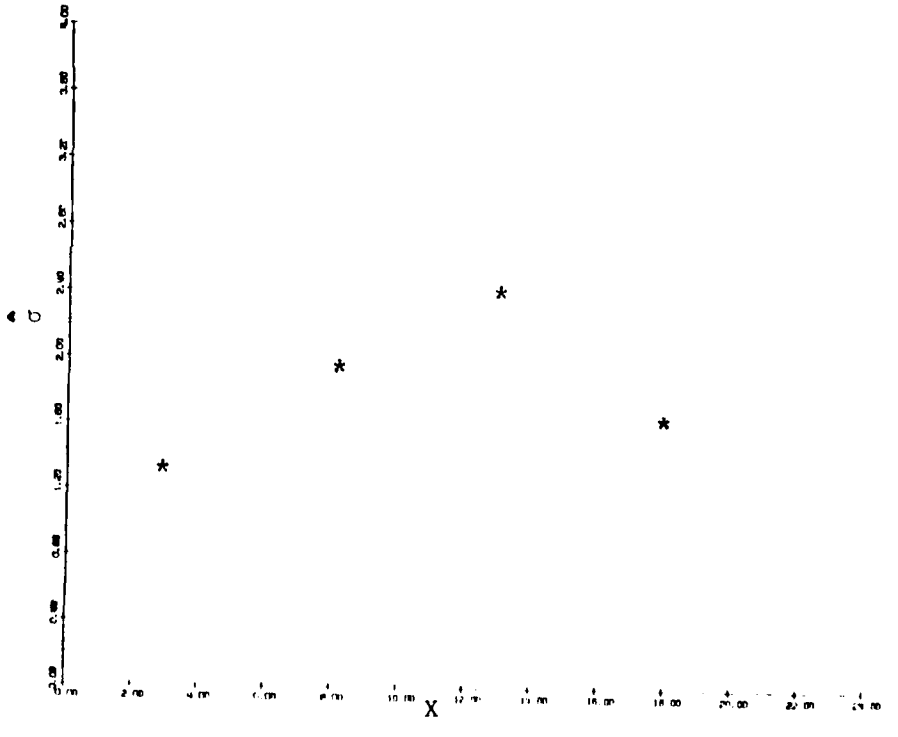
$$\mu_i = 19.8642 + .5024\ X_i$$

$$\sigma_i = 1.3914 + .0363\ X_i$$

Since we suspected that σ_i does not change significantly with X_i, we test $H_\sigma : \beta_\sigma = 0$ giving $z_\sigma = .8117$. Based on this value we fail to reject H_σ at the $\alpha = .05$ level and conclude that X_i does not have a significant linear effect on σ_i. A test of $H_\mu : \beta_\mu = 0$ gives $z_\mu = 7.943$ and we reject H_μ at the .05 level.

Figure 6.F Professors' Salary Data; Plot of $\hat{\mu}$ and $\hat{\sigma}$ versus X





Stage 4: Based on the results of Stage 3 and analogous with previous investigations of the data we estimate A(u) and B(u) for u = .25, .50, and .75. The estimated values are

$$A(.25) = 18.926 ,$$

$$B(.25) = .466 ,$$

$$A(.50) = 19.864 ,$$

$$B(.50) = .502 ,$$

$$A(.75) = 20.802 ,$$

$$B(.75) = .527 .$$

These values are comparable to those of Table 6.1.

While we reached the same general conclusions as the other investigations, our technique has several distinct advantages over the other procedures:

1) The procedure is flexible enough to incorporate virtually any specified distributional shape. Griffiths and Willcox (1978) only use the normal distribution. Angers (1979) and Hogg (1975) use nonparametric methods.
2) The procedure is computationally simple. Both Griffiths and Willcox (1978) and Angers (1979) use techniques that are fairly complicated and involve iterative solutions. Hogg's (1975) technique is graphical and seems very subjective in all phases of estimation.
3) The procedure uses simple well-known procedures for hypothesis testing.

One should note the danger of trying to predict salaries at other universities or outside the range of years of service using the results of this analysis. As the years of service increase, the quantile curves will undoubtedly flatten.

6.2 Green Sunfish Example

Matis and Wehrly (1979) illustrate compartmental modeling techniques using data from a study to investigate the resistance of the green sunfish, lepomis cyanellus, to various levels of thermal pollution. The data consist of the time until death (Y) of samples of fish exposed to water heated to a range of sub-lethal and lethal temperature (X). As part of their analysis Matis and Wehrly utilize samples at the temperature levels of 39.5°C and 39.7°C. They model the time until death as a three-parameter Weibull distribution and estimate all three parameters for each sample. LaRiccia (1979) uses the temperature levels 39.5°C, 39.6°C, and 39.7°C and using a Weibull model, he estimates all three parameters for each sample using minimum quantile distance estimators. The estimates of Matis and Wehrly (1979) and those of LaRiccia (1979) using r = 6 quantiles are summarized in Table 6.2. LaRiccia (1979) states that for the temperature levels of 39.5°C and 39.7°C the data fits well a Weibull distribution with the estimated parameters but that the estimated parameter values for a temperature of 39.6°C are unrealistic.

For this study we use the ten (k = 10) temperature levels 38.9°C, 39.0°C, 39.3°C (.1°C) 40.0°C and model each of the ten populations as a location-scale shift of a Weibull distribution with a common but unknown shape parameter. While this is a reasonable model for a time until failure distribution, it should be noted that in the ichthyological literature tolerance times of fish are often assumed to have a lognormal distribution.

Table 6.2 Estimated Parameters for Green Sunfish Data

	$\hat{\mu}$	$\hat{\sigma}$	$\hat{c} = (\frac{\hat{1}}{\gamma})$
		a. 39.5°C	
Matis and Wehrly(1979)	96. min	1.00*	3029.
LaRiccia(1979)	135.37	79.46	1.15
		b. 39.6°C	
LaRiccia(1979)	91.3	4.96×10^5	1.70×10^4
		c. 39.7°C	
Matis and Wehrly(1979)	35. min	.599*	2.486
LaRiccia(1979)	48.83	48.58	1.46

*the scale parameter Matis and Wehrly (1979) estimate is $k = (1/\sigma)^c$. The value $\hat{\sigma}$ is obtained by $\hat{\sigma} = (k)^{-1/\hat{\gamma}}$

The reasonability of our distributional assumptions is investigated in Stage 1 below. Our goals in this study are twofold:

1) to investigate if there is a significant difference in the location and scale parameters of the time until failure distributions for these temperature levels.
2) to estimate quantile regression lines for u = .50 and .90 (i.e. for the 50th and 90th percentiles of the time until failure distributions).

The sample sizes are $n_i = 20$ for $i = 1, \ldots, 7$, $n_8 = 11$, $n_9 = n_{10} = 10$. Figure 6.G presents the data plotted as a function of temperature level. The four stages of analysis are described below:

Stage 1: The quantile-box plots of all ten samples are given in Figure 6.H. The shift in location is evident but is not uniform for all the temperature levels. The decreasing spread as temperature increases is apparent by examining $\tilde{Q}_i(.75) - \tilde{Q}_i(.25)$. The plots of $\tilde{Q}_i(u)$ seem fairly symmetric except for $i = 3 (X_3 = 39.3^oC)$, $i = 6 (X_6 = 39.6^oC)$, and $i = 9 (X_9 = 39.9^oC)$. For i = 3, 6 and 8, $\tilde{Q}_i(u)$ is slightly skewed left and for i = 2 and 9, $\tilde{Q}_i(u)$ is skewed right. The plots suggest that a potential set of values of the shape parameter might be in the range (.5, .2).

Using the estimator $\tilde{\gamma}_p$ of γ defined by (3.2.6) and using $u_1 = .0002$, $u_2 = .0115$, $u_3 = .5429$ which are optimal for $\gamma = .3$, we obtain the estimate $\tilde{\gamma}_p = .416$.

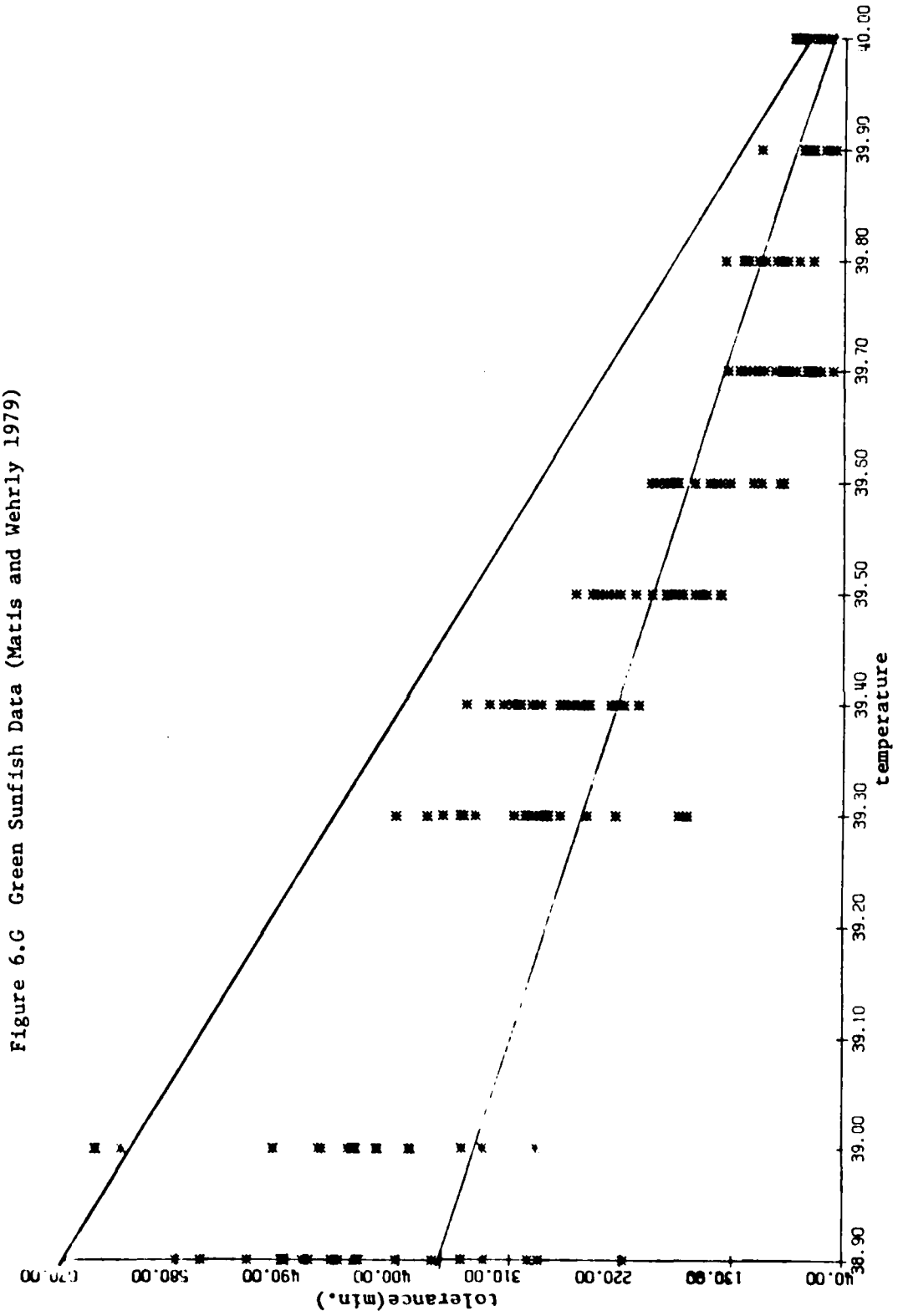


Figure 6.G Green Sunfish Data (Matis and Wehrly 1979)

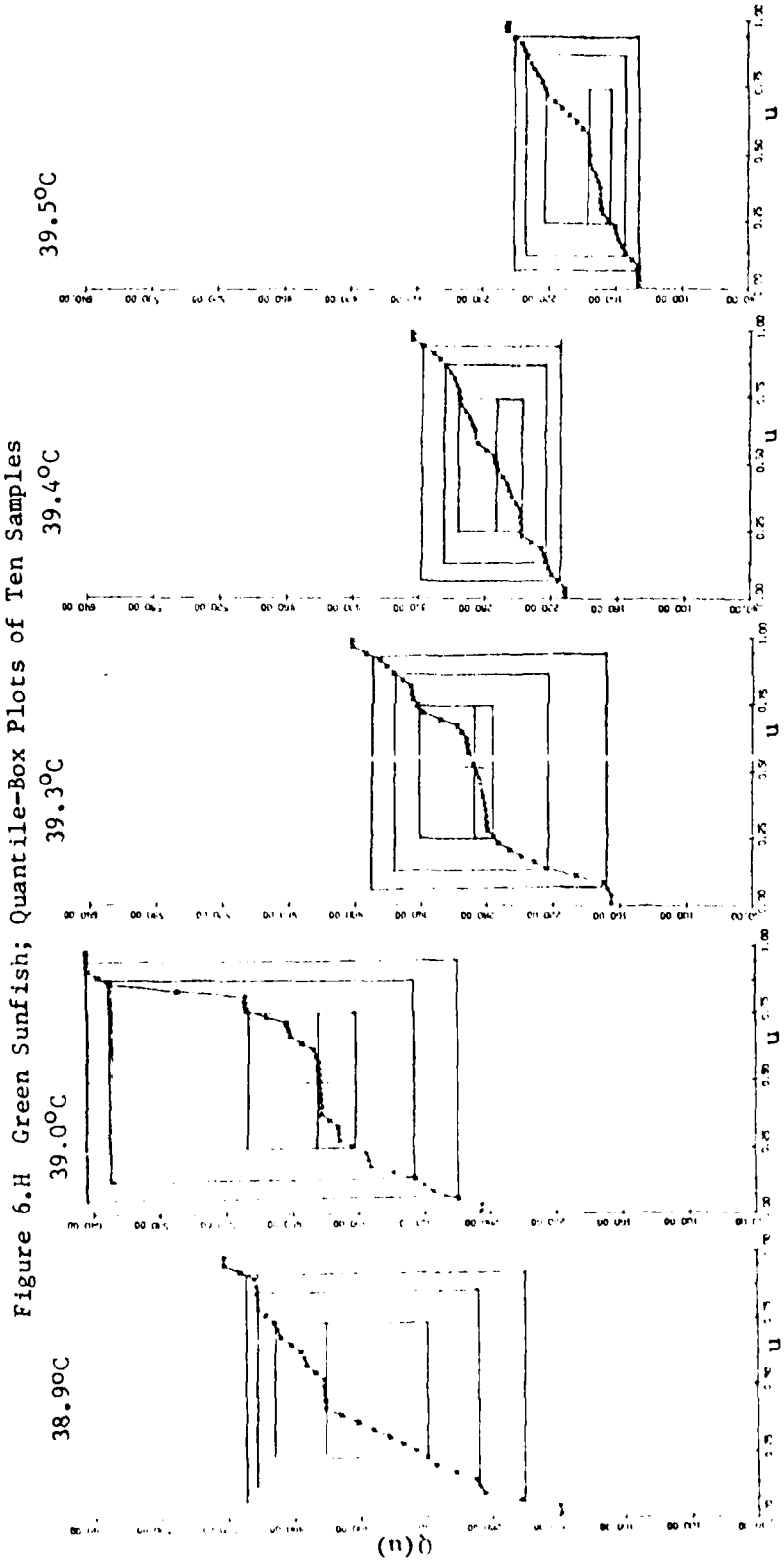


Figure 6.H Green Sunfish; Quantile-Box Plots of Ten Samples

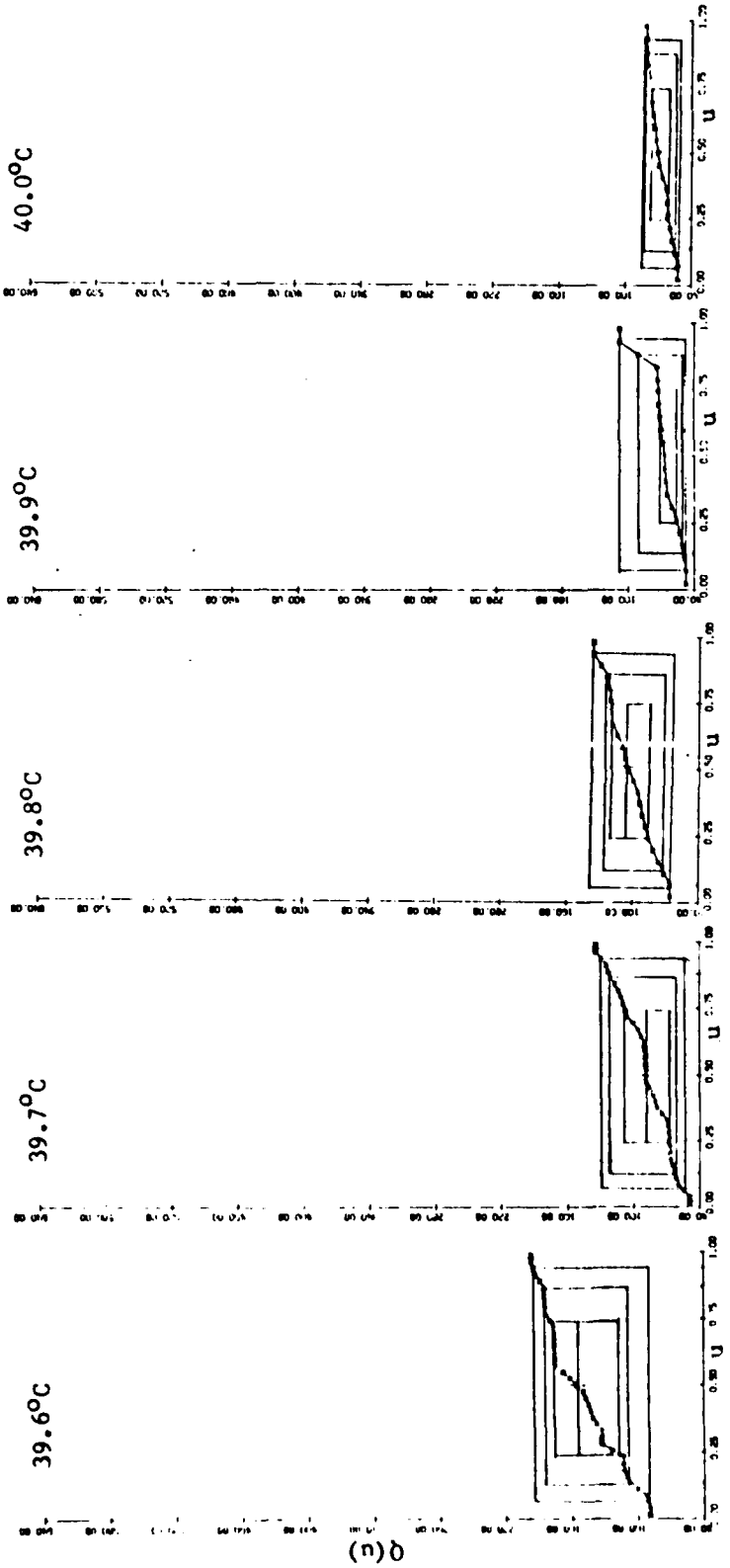


Figure 6.H (continued)

Using the technique of Section 3.3 we test the goodness-of-fit of the data to the Weibull distribution for γ_o = .333, .25, .20, .167, and .143 using the model

$$Q_i(u) = \mu_i + \sigma_i [-\log(1-u)]^{\gamma_o} , \quad i = 1, \ldots, 10 .$$

By specifying γ_o = .333, we obtain the quantile-box plot of Figure 6.I for the pooled transformed data. We are testing whether the pooled transformed data fits an exponential distribution and the plot is not incompatible with an exponential shape except for the two outliers in the right tail. Figure 6.J is a plot of $\tilde{D}(u)$, the raw transformation distribution function with the line D(u)= u superimposed on the plot. Serious departures of $\tilde{D}(u)$ from the line D(u) = u are not evident except as u gets close to 1. The value of $\tilde{\rho}(2)$ is .0097 so that comparing $2n\,\tilde{\rho}(2)$(=3.317 where n = 171) to the .05 critical value of χ^2_2 (=5.99) yields further evidence for failing to reject γ_o = .333 as the true value of γ. Finally CAT selects an optimal order of zero which is consistent with the other diagnostics in accepting γ_o = .333 as an appropriate value of γ.

The values γ_o = .25 and .20 do not prove to be acceptable values of γ based on a consensus of the diagnostics from the ONESAM analysis of the pooled transformed data. Based on this stage of the analysis we conclude that

$$Q_i(u) = \mu_i + \sigma_i [-\log(1-u)]^{.333} , \quad i = 1, \ldots, 10 .$$

Figure 6.I Green Sunfish Data; Quantile-Box Plot of Pooled Transformed Data, Weibull (γ = .333) Case

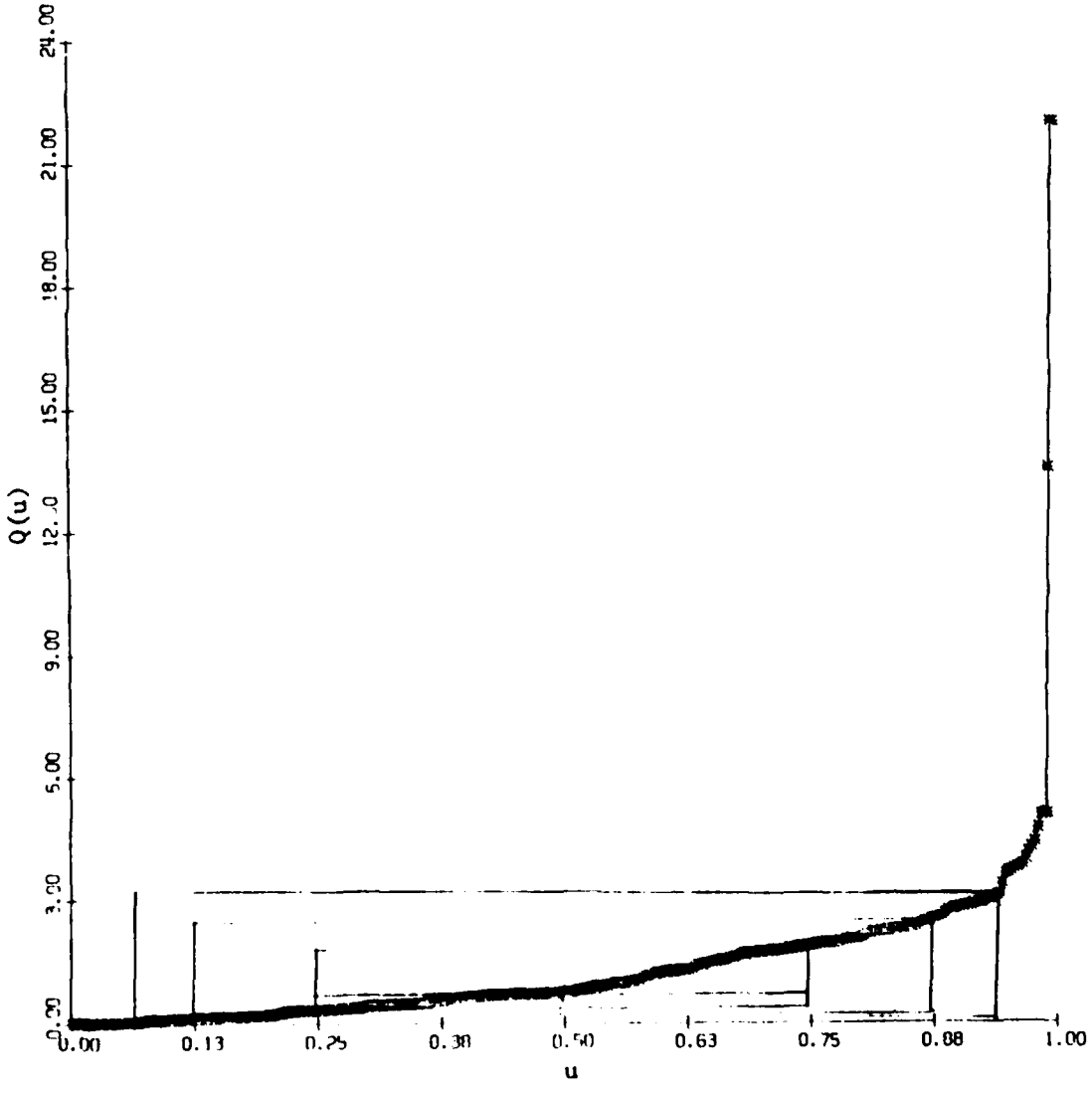
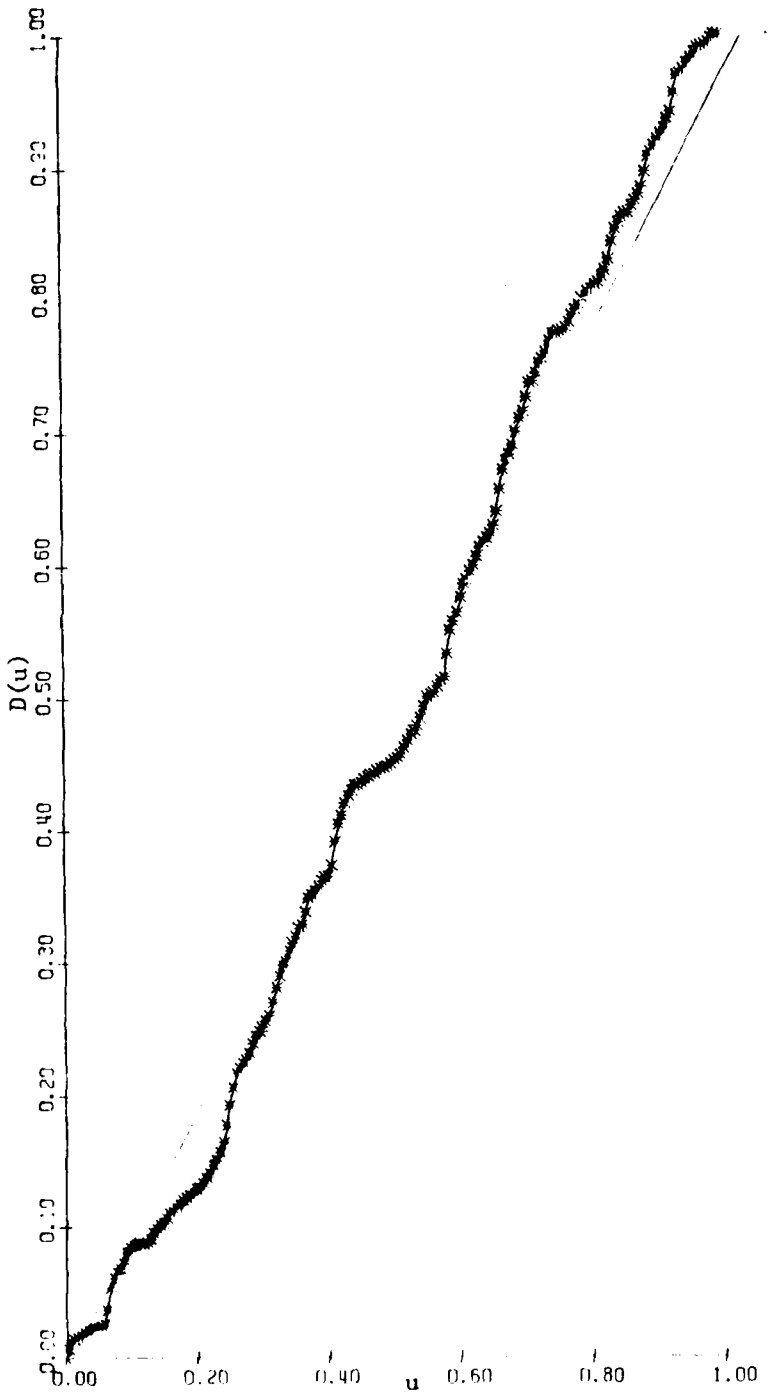


Figure 6.J Green Sunfish Data; The Function $\tilde{D}(u)$, Weibull (γ =.333) Case



Stage 2: We compute estimates $\hat{\mu}_i$, $\hat{\sigma}_i$ of μ_i, σ_i using LCOS based on the Weibull distribution with $\gamma = .333$ using the optimal coefficients and spacings ($r = 6$) of Hassanein (1971). Figure 6.K plots $\hat{\mu}_i$ and $\hat{\sigma}_i$ vs X_i. The figure suggests a linear model for $\hat{\sigma}_i$ and the presence of a linear trend for $\hat{\mu}_i$. It should be noted that $\hat{\mu}_i$ is less than min $(Y_{ij}, j = 1, \ldots, n_i)$ for all i which is desirable. However the $\hat{\mu}_i$'s vacillate so that a uniform decrease in the threshhold of the tolerance times as the temperature increases is not evident. The failure to detect a uniform trend is attributable to competing physiological causes of death in the specified temperature range.

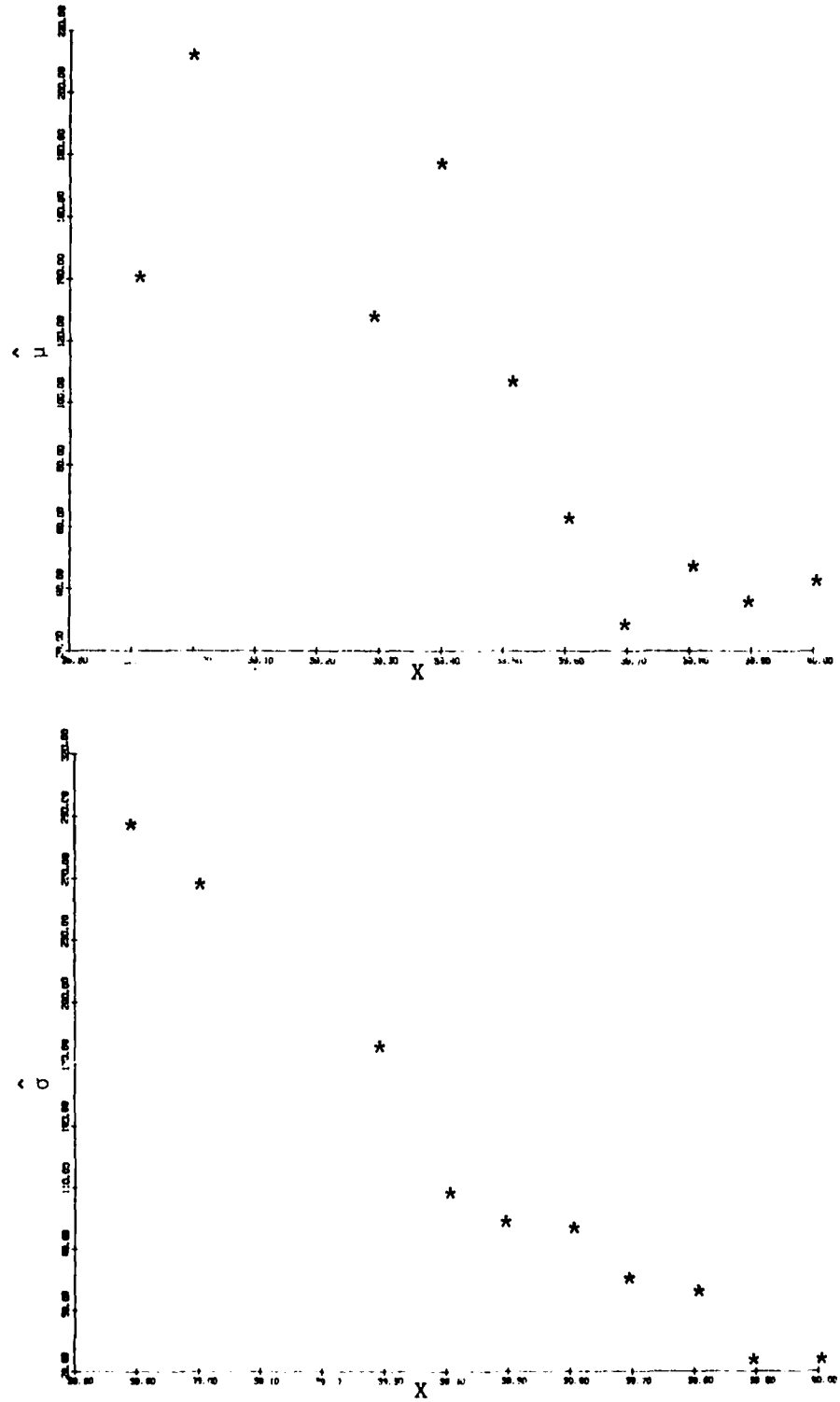
We can compare our estimates of μ, σ, and $\gamma(=1/c)$ to those of Table 6.2 for i = 5, 6, 7 (39.5°C, 39.6°C, 39.7°C). Our values are summarized in Table 6.3. The value $\gamma = .333$ which we used is not consistent with the estimate, $\hat{c}$, of LaRiccia but is consistent with that of Matis and Wehrly for the temperature 39.7°C.

Table 6.3. Estimated Parameters of Green Sunfish Data, i = 5,6,7

	$\hat{\mu}_i$	$\hat{\sigma}_i$	$\hat{\gamma}_i$	$\hat{\gamma}_p$	γ_o
i = 5, 39.5	109.319	95.899	.206	.416	.333
i = 6, 39.6	63.926	91.753	.416	.416	.333
i = 7, 39.7	29.708	66.847	.416	.416	.333

We shall attempt to fit the linear model of (5.2.1) for both μ_i and σ_i in Stage 3.

Figure 6.K Green Sunfish Data; Plot of $\hat{\mu}_i$ and $\hat{\sigma}_i$ versus X_i



Stage 3: We use generalized least squares to obtain the following fitted regression lines for μ_i and σ_i

$$\mu_i = 4601 - 114.25\ X_i$$

$$\sigma_i = 6707 - 167.188X_i\ .$$

The asymptotic variances of α_μ and α_σ are very large. One solution to this might be to rescale the X values by subtracting median (X_i) from each one. If we let $X_i^* = X_i - 39.55$, we get the regression lines

$$\mu_i = 89.560 - 124.185X_i^*$$

$$\sigma_i = 102.0931 - 181.777X_i^*$$

Testing $H_\mu: \beta_\mu = 0$, we get $z_\mu = 9.31$ and consequently we reject H_μ. Testing $H_\sigma: \beta_\sigma = 0$, we get $z_\sigma = 12.48$ and we also reject H_σ. The hypothesis H_μ is equivalent to the hypothesis that all σ_i's are equal and H_σ is equivalent to the hypothesis that all σ_i's are equal. Thus for the k sample comparison of location and scale parameters we conclude that the μ_i's are significantly different as are the σ_i's.

Stage 4: Based on the results of Stage 3 we estimate A(u) and B(u) for u = .50 and .90. The resulting quantile regression lines are

$$Q_i(.50) = 160.326 - 250.183\ X_i^*$$

$$Q_i(.90) = 324.638 - 542.742\ X_i^*$$

These lines are drawn on Figure 6.G.

The estimated quantile regression lines seem fairly reasonable. Better knowledge of the physiological effects of thermal pollution should lead to a better range of temperature levels where one effect is the dominant cause of death. Larger sample sizes will result in a better estimation of γ and of μ_i and σ_i which will improve estimates of the quantile regression line. We are convinced that this technique of quantile regression is appropriate and very useful for analyzing this type of data.

7. CONCLUSIONS

7.1 Summary:

In this dissertation we have investigated a quantile function approach to the k sample quantile regression problem. By modeling the quantile functions of the k populations as location-scale shifts of a completely specified quantile function, Q_o, and then modeling the relationship of the location and scale parameters μ_i and σ_i to a predictor variable X_i, four stages in the analysis have been delineated.

Stage 1, the identification of Q_o, is discussed in Section 3. Multiple quantile-box plots are used as a quick graphic technique to identify the qualitative characteristics, e.g. skewness, symmetry, modality, and tail behavior of the distribution of each population. Parzen's (1979) data modeling technique for one population is described and extended to a goodness-of-fit procedure for k populations. An estimator, $\tilde{\gamma}$, of the shape parameter, γ, of Q_o is given and is shown to have an asymptotic normal distribution. Optimal spacings for γ when Q_o corresponds to the Weibull distribution are given.

Section 4 describes Stage 2, the estimation of location and scale parameters using k independent samples of data. Two approaches to selecting optimal linear combinations of order statistics for one population are discussed and shown to provide computationally simple and statistically efficient estimators of the location and scale parameters of k populations. A study of bias, variance and mean

squared error of estimators based on a misspecified value of the shape parameter of the Weibull distribution shows that mild misspecification of Q_o does not seriously affect the estimation of location and scale parameters.

Stage 3, the estimation of the parameters of a linear regression model for μ_i and σ_i, is discussed in the first part of Section 5. The estimated parameters and their joint asymptotic normality are based on the generalized least squares technique. The model used for the k sample quantile regression is flexible in that it accommodates almost any specification of Q_o yet leads to simple estimators of the regression parameters.

Stage 4 is the estimation of and inference about quantile regression curves. The estimation technique is simple, and contrary to many existing techniques, one can estimate regression curves for several quantiles without having to reestimate the regression parameters. Inference about the curves is based on the asymptotic normality of the estimated parameters.

Finally in Section 6, the technique is illustrated using two data sets. In both cases an appropriate specification of Q_o is made and the estimated quantile regression curves fit the data well. The results are consistent with those of previous investigators. The two analyses illustrate the flexibility and simplicity of the k-sample quantile regression procedure.

7.2 Problems for Further Research

The most critical stage of the analysis as we perceive the k sample quantile regression problem is the identification of Q_o. There are several areas for future investigation dealing with this stage. There are a multitude of goodness-of-fit procedures for one population. The extension of these procedures to k populations and a comparison of these k population procedures to our GOF procedure should be conducted.

The estimator $\tilde{\gamma}$ (and $\tilde{\gamma}_p$) of the shape parameter γ, is formulated in general terms. Tables of optimal values for u_1, u_2, and u_3 should be available for distributions other than the Weibull, especially the lognormal distribution. The use of this type of estimator should be extended to the case of censored samples. Other methods of estimating γ, e.g. cross-validation techniques (Stone 1974), might prove useful.

Optimal linear combinations of order statistics yield estimators of location and scale parameters that are simple to compute and statistically efficient but require tables of optimal spacings and coefficients. Eubank (1979) suggests the investigation of techniques to use spacings from a subinterval of [0, 1] for distributions where the simultaneous estimation of location and scale parameters is not possible using the continuous parameter time series approach. This would be useful in the k sample quantile regression problem also. Tables of optimal spacings and coefficients

for a wider range of values of the shape parameter of the Weibull distribution need to be made available.

While this formulation of the k-sample quantile regression has proven its worth, it would be worthwhile to investigate a quantile function approach to the k-sample comparison problem.

BIBLIOGRAPHY

Angers, Claude (1979), "Simultaneous Estimation of Percentile Curves with Application to Salary Data," Journal of the American Statistical Association, 74, 621-625.

Box, G. E. P. (1954), "Some theorems on quadratic forms applied in the study of analysis of variance problems. Effects of inequality of variance in the one-way classification," Annals of Mathematical Statistics, 25, 290-302.

Brown, George W., and Mood, Alexander M. (1950), "On Median Tests for Linear Hypotheses," in Proceedings of the 2nd Berkeley Symposium, ed. J. Newman, Berkeley: University of California Press.

Chernoff, H., Gastwirth, J., and Johns, M. V. (1967), "Asymptotic Distributions of Linear Combinations of Order Statistics With Applications to Estimation," Annals of Mathematical Statistics, 38, 52-72.

Csörgö, M. and Révész, P. (1978), "Strong Approximations of the Quantile Process," Annals of Statistics, 6, 882-894.

Dubey, Satya D. (1967), "Some Percentile Estimators for Weibull Parameters," Technometrics, 9, 119-129.

Eubank, Randall L. (1979), "A Density-Quantile Function Approach to Choosing Order Statistics for the Estimation of Location and Scale," Technical Report A-10, Institute of Statistics, Texas A&M University.

Gartside, P. S. (1972), "A Study of Methods for Comparing Several Variables," Journal of the American Statistical Association, 67, 342-346.

Griffiths, David, and Willcox, Mary (1978), "Percentile Regression: A Parametric Approach," Journal of the American Statistical Association, 73, 496-498.

Harter, H. L. and Moore, A. E. (1969), "Asymptotic Variances and Covariances of Maximum Likelihood Estimators, from Censored Samples, of the Parameters of the Weibull and Gamma Populations," Annals of Mathematical Statistics, 38, 557-570.

Hassanein, Khtab (1971), "Percentile Estimators of Parameters of the Weibull Distribution," Biometrika, 58, 673-676.

——— (1972), "Simultaneous Estimation of Parameters of the Extreme Value Distribution by Sample Quantiles," Technometrics, 14, 63-70.

Hogg, Robert V. (1975), "Estimates of Percentile Regression Lines Using Salary Data," Journal of the American Statistical Association, 70, 56-59.

Johnson, Norman and Kotz, Samuel (1970), Continuous Univariate Distributions-1, Boston: Houghton Mifflin Co.

Kübler, Heiner (1979), "On the Fitting of the Three-Parameter Distributions Lognormal, Gamma, and Weibull," Statistichen Hefte, 20, 68-125.

LaRiccia, Vincent N. (1979), "A Family of Minimum Quantile Distance Estimators," unpublished Ph.D. dissertation, Institute of Statistics, Texas A&M University.

Matis, J. H., and Wehrly, T. E. (1979), "Stochastic Models of Compartmental Systems," Biometrics, 35, 199-220.

Ogawa, J. (1951), "Contributions to the Theory of Systematic Statistics, I," Osaka Mathematical Journal, 3, 131-142.

Parzen, Emanuel (1979a), "Nonparametric Statistical Data Modeling," Journal of the American Statistical Association, 74, 105-121.

——— (1979b), "A Density-Quantile Function Perspective on Robust Estimation," Robustness in Statistics, ed. by Robert Larmer and Graham Wilkinson, New York: Academic Press.

——— (1980), "Data Modeling Using Quantile and Density-quantile Functions," Technical Report B-2, Institute of Statistics, Texas A&M University.

———, and White, J. Michael (1979), "ONESAM, A Computer Program for Nonparametric Data Analysis and Goodness of Fit," Technical Report A-7, Institute of Statistics, Texas A&M University.

Rao, C. Radhakrishna (1973), Linear Statistical Inference and Its Applications, New York: John Wiley and Sons.

Stigler, S. (1974), "Linear Functions of Order Statistics With Smooth Weight Functions," Annals of Statistics, 2, 676-693.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," Journal of the Royal Statistical Society-B, 36, 111-147.

Tsai, W. S., Duran, B. S., and Lewis, T. O. (1975), "Small Sample Behavior of Some Multisample Nonparametric Tests for Scale," Journal of the American Statistical Association, 70, 791-796.

Tukey, John W. (1977), Exploratory Data Analysis, Reading, Mass: Addison-Wesley Publishing Co.

Zanakis, Stelios H. (1979a), "A Simulation Study of Some Simple Estimators for the Three-Parameter Weibull Distribution," Journal of Statistical Computing and Simulation, 9, 101-116.

——— (1979b), "Extended Pattern Search with Transformations for the Three-Parameter Weibull MLE Problem," Management Science, 25, 1149-1161.


ATE